Data consistency in summary measures of population health

Michelle E. Kruijshaar

Data consistency in summary measures of population health,
Kruijshaar, Michelle E.
Thesis Erasmus MC, University Medical Center Rotterdam,
with summary in English and Dutch.

Cover design: Bon Mot, Rotterdam
Layout: Michelle E. Kruijshaar
Pinted by: PrintPartners Ipskamp, Enschede

ISBN 90-9017657-8

Data consistency in summary measures of population health

Data consistentie in samengestelde volksgezondheidsmaten

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof. Dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 21 januari 2004 om 13:45 uur
door

Michelle Elisabeth Kruijshaar

geboren te Amsterdam

Promotiecommissie

Promotoren:	Prof. dr. P. J. van der Maas
	Prof. dr. ir. D. Kromhout
Overige leden:	Prof. dr. F. C. Verhulst
	Prof. dr. G. J. Bonsel
	Prof. dr. A. Hofman
Copromotoren:	Dr. J.J. Barendregt
	Dr. N. Hoeymans

Contents

1

General introduction

Introduction

Summary measures of population health combine information on different aspects of health to depict a population's health as a single figure. They may be used to identify differences in health (e.g. between groups or countries), or changes over time (e.g. as a result of policy measures) [1]. The interest in and use of these measures has increased over the past decades, and different types have been constructed [2]. Although it is theoretically attractive to include information on specific diseases in summary measures (for reasons that will be discussed later in this chapter), in practice data problems limit the use of disease-specific summary measures. In this thesis we study practical data problems that may restrict the construction of these measures. In the introductory chapter we first define summary measures of population health and distinguish different types. Then we describe the data problems and the research question they gave rise to. We end this chapter with an outline of the thesis.

Measuring population health

Health policy makers are concerned with protecting and improving the health of a population. To inform decision-making and to evaluate past policy decisions, quantitative information is needed on the level of and changes in a population's health. Measures that describe the level of health of a population are referred to as "indicators of population health". These indicators can cover different aspects of health. Some reflect quantitative aspects of the occurrence of fatal and non-fatal health outcomes, such as mortality rates, life expectancy, incidence and prevalence of diseases. Others quantify aspects concerning the quality of life of individuals and their ability or disability to function. Health-related quality of life measures and health status evaluations provide such information.

Health indicators, when adopted to inform health-policy making and target setting, acquire a normative aspect because health policy focuses on improving health outcomes that are measured by the health indicators used [3]. In the past, decreasing infant and general mortality was the major challenges for health policy. Most health indicators used at that time reflected this focus on mortality (e.g. the infant mortality rate and life expectancy at birth) and policy targets were formulated in terms of these indicators. As a consequence, health policy focused mainly on preventing premature death. However, after the large

gains in life expectancy made in the first half of the past century, improving quality of life became increasingly important. With longevity achieved for many people, the aim of health policy shifted towards improving both quality and quantity of life. Since both quality and quantity of life are important, both should be incorporated in indicators of population health and indicators based only on mortality are no longer sufficient [4, 5]. For this reason, summary measures of population health were constructed.

Summary measures of population health

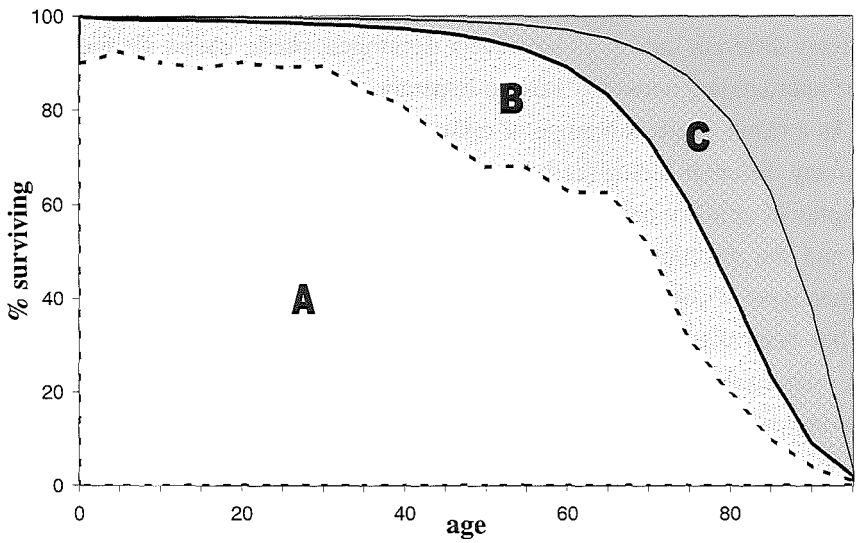
Summary measures of population health integrate data on mortality with information on non-fatal health outcomes to represent the health of a population as a single number [6]. These two types of data are combined using time as a common denominator. Time spent in optimal health and time spent in sub-optimal health can be compared using information on how people value time lived in different health states. Thus, not only age-specific data on the epidemiology of fatal and non-fatal health outcomes are required as basic input to all summary measures, but also information on valuations attached to health states.

Since the first efforts were made to construct this kind of measures [7-10] many different indicators have been developed, partly within the framework of the international network "Réseau Espérance de Vie en Santé" (REVES) [11, 12]. As a result, there is a range of options in the design of summary measures, and many different indicators have been used in different places. For example, examples of applications of measures such as the active life expectancy (ALE) [13] and the disability-free life expectancy (DFLE) [14, 15] are widespread. Another type of measure, the Disability Adjusted Life Years (DALY) [3, 16, 17], has been used in the Global Burden of Disease Study (GBD) [3] and in many subsequent national burden of disease studies (e.g. [18-21]). In the following two sections we will make two broad distinctions: between health expectancies and health gap measures, and between generic and disease-specific summary measures. This will divide summary measures into four broad groups. Further diversification, which will not be discussed here, is possible by variations in the calculation method; by using different definitions, measurement and valuation of health states; and by introducing additional values into the indicator (e.g. age-weights and discount rates) [2, 22].

Health expectancy and health gap measures

Summary measures can be divided into two broad families on the basis of a simple survivorship function: health expectancies and health gaps [2]. In Figure 1.1 the middle curve shows the survival curve for males in the Netherlands in 1999. For each age it reflects the percentage of men from an initial birth cohort that has survived up to that age. The area under this curve (Area A + B) represents the male life expectancy at birth in the Netherlands. Health expectancies are functions of this area that take into account that a proportion of the time lived was spent in less than optimal health (area B). They add the life years lived in optimal health (area A) and those lived in less than optimal health (area B), giving a lower weight to the latter. The weights assigned to years in area B range between 0 and 1, with 1 reflecting optimal health and 0 being equivalent to years lost to premature mortality. A well-known example of a

Figure 1.1. **Survivorship curve for men in the Netherlands in 1999.**



Health Expectancies (HE) and Health Gaps (HG) are calculated as follows: HE = A + f(B), and HG = C + g(B), in which the functions f and g assign a weight between 0 and 1 to years lived in less than optimal health, with optimal health being represented by 1 in function f and by 0 in g, and time lost to premature mortality by 0 and 1, respectively.

health expectancy measure is the Disability-Free Life Expectancy (DFLE). The DFLE combines general life expectancy with information on the prevalence of disability within a population (DFLE) [14, 23], assigning all years lived with disability (above a certain level) a weight of 0. Other examples are the active life expectancy (ALE) [13], the years of healthy life (YHL) [24], and the dementia-free life expectancy [25].

Health gaps on the other hand measure the difference between the actual health of a population and some stated norm. For example, in Figure 1.1 the proposed norm is reflected by the upper curve, the 75% lowered mortality curve. The mortality gap in Figure 1.1 is then represented as area C. Health gaps, apart from measuring the mortality gap, can also take into account the fact that not all years lived were spent in optimal health. Health gaps are thus functions of the areas B and C. They add the years of life lost (area C) and the years of life lived in less than optimal health (area B) counting the latter only partially. The weight assigned to the years in area B again range between 0 and 1, but with 1 now being equivalent to time lost to premature mortality and 0 representing optimal health. An example of a health gap measure is the Disability Adjusted Life Year (DALY), that was developed within the Global Burden of Disease Study 1990 (GBD 1990) [3, 16, 17].

Generic and disease-specific summary measures

Depending on the epidemiological input used to construct summary measures, we distinguish generic and disease-specific measures. Generic measures are based on generic information on mortality, morbidity and health status. An example is the DFLE that combines generic life expectancy with generic information on disability. All examples of generic indicators (e.g. DFLE, ALE, YHL) are health expectancy measures: no generic gap measure has (yet) been constructed. The other group of measures, the disease-specific measures, uses disease-specific data as input. In this group examples of both health expectancies and health gaps can be found. The Disability Adjusted Life Expectancy (DALE) [26-28], that uses the inverse health status values of the DALY, is an example of a health expectancy measure. The dementia-free life expectancy is another. The best-known example in this group, however, is a gap measure: the DALY. Measures can be specific for just one disorder, like the dementia-free life expectancy, but can also include information for a group or cluster of diseases,

as was for example done in the GBD 1990 for the DALY [3, 16, 17] and by Barendregt et al. for the DALE [26].

The choice between a generic and a disease-specific measure depends on its intended use. Generic health expectancy measures have been calculated for many countries [29] and provide insight into questions such as: "Are the life-years gained lived in good health?". Generic measures provide a tool to monitor the overall health of a population and, at least theoretically, to detect changes therein and compare countries. However, this information is useful to guide health policy decisions only if changes or differences in health can be attributed to certain diseases or determinants [26]. Policy makers face the difficult task to allocate resources and to choose between different intervention and research programmes that are mostly disease- or risk factor-oriented. Therefore, disease-specific information (or disease-cluster-specific) is more useful to inform decision-makers.

The best known disease-specific measure, the DALY, describes the burden that a specific disease causes within a population by combining disease-specific epidemiological frequency data on morbidity and mortality and disease-specific health status values. For each disease, it is calculated adding the number of Years of Life Lost due to premature mortality (YLL) and the number of Years Lived with Disability (YLD) [3]. The latter, the years lived with a specific disease, are weighted for the severity of the disability that is associated with the disease using a set of disease-specific disability weights (DWs). These weights are indispensable, as they make the years lived with different diseases comparable, enabling the burden of different diseases to be added and to be compared (enabling ranking). Ultimately, the addition of the disease-specific burdens for all diseases results in a generic estimate of the overall population burden. This method, used in the GBD 1990 [3] to estimate the overall population burden, is referred to as a bottom-up approach: a generic measure is built up from its disease-specific parts. The opposite, the top-down approach, refers to a generic population measure being broken down into disease-specific components. An example of this is the decomposition of the survival curve into underlying causes of death [30].

Data problems of disease-specific summary measures

From a health policy perspective disease-specific summary measures of population health thus have an advantage over generic measures. However, in practice, their use may be limited due to their large data requirements. Disease-specific measures require, for all diseases, disease-specific epidemiological frequency data on morbidity and mortality, as well as disease-specific health status valuations. Both types of data bring about specific problems to the use of disease-specific measures.

The epidemiological frequency data are not always available and, if they are, can be incomplete and/ or of questionable validity. In the Netherlands disease-specific prevalence and/ or incidence data are not collected for every disease, and when they are, their validity can be affected by selection bias, unclear case-definitions, problems with self-reporting and other problems. Cause-specific mortality data on the other hand are available in the Netherlands (from the cause-of-death registration), but their validity is never perfect as long as ambiguity in the primary cause of death exists.

To remedy some of these problems, models have been developed that exploit the causal relationship between the different epidemiological parameters for one disease. An example are IPM models (incidence-prevalence-mortality models). They exploit the fact that in the causal pathway of every disease incidence has to precede prevalence, and disease-specific mortality can only follow from having the disease. IPM models have been used both to infer missing data from the available information and to check for the internal consistency of data. Inconsistencies may be caused by differences in the completeness of the data. For example, when more incident cases are missed than deaths, incidence and mortality are inconsistent. Also, they may arise when the data were derived from different contexts (e.g. in different regions) or were measured differently (e.g. with varying case-definitions). Applying inconsistent data to an IPM model results in discrepancies between model outcome and data. In this way inconsistencies can be detected and the observed data adjusted for the inconsistencies. Furthermore, when one parameter is missing, IPM models allow it to be calculated from the existing data. The validity and usefulness of these models, however, has not been studied yet.

The key issue of the health status valuations (disability weights in the DALY methodology) is that it has to be ensured that they refer to the same

health status for which the epidemiological data are collected. They should be tailored to the epidemiological case-definition, to guarantee that both refer to exactly the same severity level of the disease. This is not always straightforward, especially for heterogeneous disorders and diseases with case-definition problems. In psychiatric epidemiology, for example, different screening instruments use different cut-off points [31], with some including more mild cases and others being more stringent. Combining prevalence estimates based on a strict instrument with a health status value based on a mild disease description will inevitably lead to an underestimation of the problem, and vice versa.

In the original valuation method of the GBD 1990 disability weights (DWs) were derived for a wide range of conditions, but the disease descriptions were often not very detailed, making it difficult to tailor them properly to the epidemiology. In a Dutch study, the GBD-1990 valuation method was further refined [32, 33]. Diseases were divided into disease stages that were more homogeneous regarding disability and treatment, for example different severity classes (mild/ severe) or stages in the disease pathway (initial stage/ end stage). Furthermore, a standardised health status description was provided for each stage using the EuroQol 5D+C system [34]. This system describes six dimensions of health (mobility, self-care, usual activities, pain/ discomfort, anxiety/ depression and cognition) in three levels of severity of the problems. The distinction of disease stages and the addition of a formal description facilitate the tailoring of the DWs to the epidemiology [35]. However, the question then becomes how to combine the disease-stage-specific DWs into one disease-specific weight.

Apart from these two data-related problems, several other problems also complicate the calculation of disease-specific summary measures. Difficulties arise in particular when disease-specific DALYs are combined to estimate the overall burden of disease. One problem is caused by co-morbidity of diseases. The health status value of having two diseases at the same time is not simply the sum of the values associated with each condition separately. For example, co-morbidity of arthritis and visual problems in older people may exacerbate the disability of the separate conditions [36]. Co-morbidity can lead to both higher and lower health status values than the sum of the separate values. This should be accounted for when the burdens for separate diseases are combined. Another problem is that in a bottom-up approach disease-specific measures can not cover all morbidity that is present in a population. The number of diseases is large,

making it difficult to be exhaustive. Consequently, rare diseases with little impact will often be ignored. Furthermore, a decrease in health status can not always be linked to a specific disease. Ageing for example may cause a decrease in functioning, and some causes of morbidity we may simply not know (yet). Adding only the disease-specific components will thus underestimate the total burden of disability. Both reasons cause disease-specific summary measures to ignore a part of the morbidity that is present in a population: the rest-morbidity.

The research questions

As health policy decisions are usually taken at the level of a disease or risk factor, there are strong theoretical arguments for the use of disease-specific measures. However, the data and methodological problems form clear restrictions to their application. The overall research question of this thesis is whether useful disease-specific summary measures can be constructed in practice. In this context, we focus on the two data problems, because they form the first bottleneck in the use of these measures. Before it is possible to study any other issues related to disease-specific health indicators, it has to be practically feasible to calculate them from the data, making the problems of co-morbidity and rest-morbidity of secondary importance. Two specific questions are addressed:

1) To what extent are causal disease models valid and useful to check the consistency of epidemiological frequency data and to supplement them?
2) How can disease-stage specific health status valuations be tailored to epidemiological frequency data?

Outline of this thesis

We explored these two research questions on the basis of empirical data for two common diseases in the Netherlands: breast cancer and major depression. We chose these disorders for two reasons: they are both important health problems and they allow us to study the research questions from different perspectives. Breast cancer is important in particular because of the mortality it causes, while depression causes mainly morbidity. Also, for breast cancer epidemiological data are easily available and regarded as relatively reliable, whereas for major depression the data still suffer from several problems. Finally, the disease staging for breast cancer and major depression are based on different concepts.

For major depression, stages were differentiated according to severity classes (e.g. mild, severe), while for breast cancer phases in the disease pathway were used (e.g. diagnosis and therapy, metastasised).

The research in this thesis consists of two parts. In part A we address the first research question: the validity and usefulness of disease models. This question can best be studied using data for a disease with well-described epidemiology. As cancer incidence and mortality are registered on a regular basis in the Netherlands and are regarded as relatively reliable, such data provide a good basis for studying this question. Chapter two therefore studies the validity and usefulness of IPM models using relatively reliable and complete data sets on breast cancer and three other common types of cancer. The results of these analyses showed us that time-trends in the epidemiological frequency data bias the outcome of these models. For breast cancer many additional epidemiological data (e.g. survival, prevalence, etc.) are available, allowing us to quantify, in chapter three, the impact of data problems and trends on the model for breast cancer. The last chapter of part A, chapter four, describes an application of a disease model to the less well monitored epidemiology of major depression to obtain internally consistent estimates for the epidemiological parameters of major depression.

The second part of this thesis, part B, is concerned with tailoring health status valuations to the epidemiology and assessing their impact on the resulting summary measure. Since tailoring is a problem especially in diseases that are heterogeneous and/ or have unclear case-definitions, we thought it relevant to study this problem for major depression. In the Netherlands, the Netherlands Mental Health Survey and Incidence Study (NEMESIS) provided a good database for major depression with information on both prevalence and health status by severity class. These data enabled us to use the severity classes in the tailoring of the DWs to the epidemiology. In chapter five we compare disability between the severity classes. Chapter six uses this information to derive health status values per severity class that we subsequently used in a burden of major depression calculation. A comparison of the results with studies using non-tailored values gives an impression of the importance of health status values on the overall burden of disease calculation. For breast cancer the health status valuations can be tailored using a modelling approach. This approach is used in chapter seven to calculate and compare the burden of breast cancer in six

European countries and to study its sensitivity to variations in health status values.

Chapter eight, the general discussion, integrates and discusses the results from these studies.

References

1. Maas PJ, van der. Applications of summary measures of population health. In: Murray CJL, Salomon J, Mathers CD, Lopez AD, editors. *Summary measures of population health: concepts, ethics, measurement and applications*. Geneva: WHO; 2002. p. 53-60.
2. Murray CJ, Salomon JA, Mathers C. A critical examination of summary measures of population health. *Bull World Health Organ* 2000;78(8):981-94.
3. Murray CJL, Lopez AD, editors. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge, MA: Harvard School of Public Health; 1996.
4. Mathers CD, Salomon JA, Murray CJ. Infant mortality is not an adequate summary measure of population health. *J Epidemiol Community Health* 2003;57(5):319.
5. Robine JM. The relevance of population health indicators. *J Epidemiol Community Health* 2003;57(5):318.
6. Field MJ, Gold MR, editors. *Summarizing population health: directions of the development and application of population metrics.* Washington, D.C.: National Academy Press; 1998.
7. Fanshel S, Bush JW. A health-status index and its application to health services outcomes. *Operations Research* 1970;18(6):1021-65.
8. Chiang CL. An index of health: mathematical models. Washington, DC.: National Center for Health Statistics; 1965. Report No.: Vital and health statistics series 2, no 5.
9. Sanders BS. Measuring community health levels. *Am J Public Health* 1964;54(7):1063-70.
10. Sullivan DF. A single index of mortality and morbidity. *HSMHA Health Rep* 1971;86(4):347-54.
11. Robine JM, Mathers CD, Bone MR, Romieu I, editors. *Calculation of health expectancies: harmonization, consensus achieved and future perspectives. Proceedings of the 6th meeting of the international network on health expectancy and the disability process REVES.* Paris: John Libbey Eurotext; 1993.

12. Mathers CD, McCallum J, Robine JM, editors. *Advances in health expectancies. Proceedings of the 7th meeting of the international network on health expectancy and the disability process REVES.* Canberra: Australian Institute of Welfare; 1994.
13. Katz S, Branch LG, Branson MH, Papsidero JA, Beck JC, Greer DS. Active life expectancy. *N Engl J Med* 1983;309(20):1218-24.
14. Robine JM, Ritchie K. Healthy life expectancy: evaluation of global indicator of change in population health. *BMJ* 1991;302(6774):457-60.
15. Mathers CD. Health expectancies: an overview and critical appraisal. In: Murray CJL, Salomon J, Mathers CD, Lopez AD, editors. *Summary measures of population health: concepts, ethics, measurement and applications*. Geneva: WHO; 2002. p. 177-9.
16. Murray CJ. Quantifying the burden of disease: the technical basis for disability- adjusted life years. *Bull World Health Organ* 1994;72(3):429-45.
17. Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study [see comments]. *Lancet* 1997;349(9063):1436-42.
18. Bowie C, Beck S, Bevan G, Raftery J, Silverton F, Stevens A. Estimating the burden of disease in an English region. *J Public Health Med* 1997;19(1):87-92.
19. Melse JM, Essink-Bot ML, Kramers PG, Hoeymans N. A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *Am J Public Health* 2000;90(8):1241-7.
20. Mathers CD, Vos ET, Stevenson CE, Begg SJ. The Australian Burden of Disease Study: measuring the loss of health from diseases, injuries and risk factors. *Med J Aust* 2000;172(12):592-6.
21. Peterson S, Backlund I, Diderichsen F. Burden of disease in Sweden - a Swedish DALY calculation. [In Swedish.]. Stockholm: National Public Health Institute [Folkhälsoinstitutet]; 1998.
22. Maas PJ, Kramers PGN, editors. *Volksgezondheid Toekomst Verkenning 1997.III Gezondheid en levensverwachting gewogen.* Maarssen: Elsevier/ De Tijdstroom; 1997.
23. Robine JM, Brouard N, Colvez A. Indicators of disability-free life expectancy. Global indicators of the health status of populations. [in French]. *Rev Epidemiol Sante Publique* 1987;35(3-4):206-24.
24. Erickson P, Wilson R, Shannon I. Years of healthy life. *Healthy People 2000 Stat Notes* 1995(7):1-15.
25. Ritchie K, Robine JM, Letenneur L, Dartigues JF. Dementia-free life expectancy in France. *Am J Public Health* 1994;84(2):232-6.
26. Barendregt JJ, Bonneux L, Maas PJ, van der. Health Expectancy: from a population health indicator to a tool for policy making. *Journal of Aging and Health* 1998;10(2):242-258.

27. Mathers CD, Sadana R, Salomon JA, Murray CJ, Lopez AD. Healthy life expectancy in 191 countries, 1999. *Lancet* 2001;357(9269):1685-91.
28. Murray CJ, Lopez AD. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global Burden of Disease Study. *Lancet* 1997;349(9062):1347-52.
29. Mathers CD. Health expectancies in the different regions of the world. Introduction. In: Robine JM, Jagger C, Mathers CD, Crimmins EM, Suzman RM, editors. *Determining Health Expectancies*. Chichester: Wiley; 2003. p. 285-8.
30. Nusselder WJ, Mackenbach JP. Rectangularization of the survival curve in The Netherlands: an analysis of underlying causes of death. *J Gerontol B Psychol Sci Soc Sci* 1997;52(3):S145-54.
31. Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001;31(6):1001-13.
32. Stouthard MEA, Essink-Bot ML, Bonsel GJ, Barendregt JJ, Kramers PGN, van de Water HPA, et al. Disability weights for diseases in the Netherlands. Rotterdam, the Netherlands: Department of Public Health, Erasmus University Rotterdam; 1997.
33. Stouthard MEA, Essink-Bot ML, Bonsel GJ, on behalf of DDW Group. Disability weights for diseases in the Netherlands. A modified protocol and results for a Western European region. *Eur J Public Health* 2000;10:24-30.
34. Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol* 1999;52(4):293-301.
35. Mathers CD, Vos T, Stevenson C. The burden of disease and injury in Australia. Canberra: AIHW; 1999. Report No.: PHE 17.
36. Verbrugge LM, Lepkowski JM, Imanaka Y. Comorbidity and its impact on disability. *Milbank Q* 1989;67(3-4):450-84.

PART A

The validity and usefulness of disease models

2

The use of models in the estimation of disease epidemiology

Michelle E. Kruijshaar, Jan J. Barendregt, Nancy Hoeymans. The use of models in the estimation of disease epidemiology. *Bulletin of the World Health Organization* 2002; 80 (8): 622-628.

Summary

Objective

To explore the usefulness of incidence–prevalence–mortality (IPM) models in improving estimates of disease epidemiology.

Methods

Two artificial and four empirical data sets (for breast, prostate, colorectal, and stomach cancer) were employed in IPM
models.

Findings

The internally consistent artificial data sets could be reproduced virtually identically by the models. Our estimates often differed considerably from the empirical data sets, especially for breast and prostate cancer and for older ages. Only for stomach cancer did the estimates approximate to the data, except at older ages.

Conclusion

There is evidence that the discrepancies between model estimates and observations are caused both by data inaccuracies and past trends in incidence or mortality. Because IPM models cannot distinguish these effects, their use in improving disease estimates becomes complicated. Expert opinion is indispensable in assessing whether the use of these models improves data quality or, inappropriately, removes the effect of trends.

Introduction

Quantitative descriptions of disease epidemiology, such as incidence, prevalence and mortality, by age and sex, are essential inputs for burden of disease studies and cost-effectiveness analyses of interventions. Such studies serve as an important source of information for policy-making, planning, and research prioritization in health care. Empirical observation is obviously the gold standard for obtaining epidemiological information, but empirical data are often incomplete or of dubious validity. In addition, the validity of estimates tends to vary even for an individual disease. For example, in instances where incidence is more difficult to observe than mortality, more incident cases than deaths are likely to be missed. In this case, therefore, data on incidence are less complete than those on mortality, making these two parameters internally inconsistent.

One way to circumvent these data limitations is to exploit the causal structure of the disease process: incidence has to precede prevalence, and cause-specific mortality can only follow being diseased. Incorporating the causal structure into a mathematical model makes it possible to calculate data that are missing from the observational set and to check for the internal consistency of observations. An example of the first of these procedures is the back-calculation of (unobserved) human immunodeficiency virus (HIV) infection from data on the incidence of acquired immunodeficiency syndrome (AIDS) [1].

The Global Burden of Disease 1990 study, and many subsequent national burden of disease studies, made extensive use of DisMod, a generic mathematical disease model, which was specially designed to supplement observational data and check for internal consistency [2]. Previously we have employed a conceptually similar disease model to calculate unobserved incidence data [3]. The present article explores the usefulness of such generic disease models that describe the relation between incidence, prevalence and mortality (IPM models) for improving estimates of disease epidemiology. We consider whether these models calculate the correct results and, if so, how useful these results are. Our approach is to apply two IPM models (DisMod and our own model) to two artificial data sets known to be complete and consistent and to four high-quality empirical data sets for cancers, drawn from Dutch registries. The ability of the IPM models to describe adequately the data sets serves as an indicator of their usefulness.

Methods

Artificial data sets

In order to demonstrate that the models can calculate the correct results, we first used them with internally consistent data (formal validity). Data sets for breast and colorectal cancer were generated by MISCAN, a microsimulation model for the evaluation of screening programmes [4, 5]. MISCAN creates a cohort of hypothetical individuals, each of whom has a risk of developing cancer, and, once the disease is present, a survival drawn from a lognormal distribution. Incidence, prevalence and mortality data generated by this model are, by definition, complete and internally consistent.

Empirical data

We applied the IPM models to national incidence and mortality data for breast cancer (ICD-9 code 174), prostate cancer (ICD-9 code 185), colorectal cancer (ICD-9 codes 153 and 154) and stomach cancer (ICD-9 code 151). Data averaged for 1991–95, specified by sex and 5-year age group (up to ≥95 years), were used. Statistics Netherlands (CBS) collects mortality data by cause-of-death on a continuous basis using information from death certificates. Incidence data are collected continuously by the Dutch Cancer Registry (NKR), which receives its information from nine regional cancer centres. These data are based on pathology reports, complemented by national hospital admission data; death certificates are not used as an additional source.

The cancer registries do not estimate prevalence data on a regular basis. The Regional Cancer Centre South (IKZ) determined the prevalence of the specific cancers for which incidence has been determined for the eastern part of the coverage area on 1 January 1993: for all incident cases registered in the region from 1970 until 1992 the population registry was checked to determine whether the persons concerned were still alive. For the same region we obtained the regional mortality and incidence rates from IKZ, averaged for 1991–95. Mortality data for this region originated from the CBS database (region: COROP 36 and 37). Regional data were specified by sex and 5-year age group (up to ≥85 years).

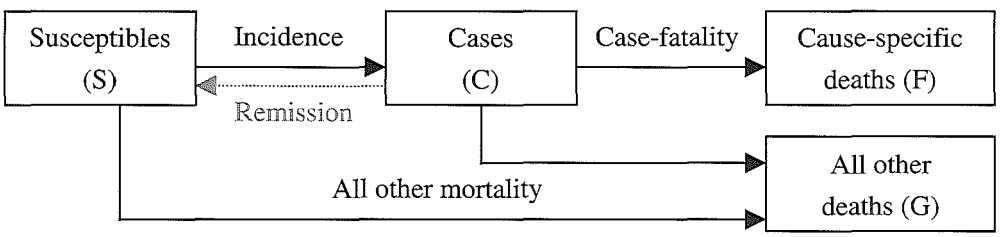
IPM models

Both DisMod and our model are based on the conceptual disease model depicted in Figure 2.1.The population is described as being in different states, while

transition hazards determine how people move from one state to another. Within a population, individuals can be either susceptibles or cases. Cases may die from their disease, while both cases and susceptibles are at risk of dying from other causes. There are consequently three transition hazards: incidence, case-fatality, and all other mortality. DisMod also includes remission as a fourth transition hazard, but for our analysis we set this hazard to zero since cure is not taken into account in the registered cancer prevalence. The framework in Figure 2.1 shows that the number of cases can be calculated by following an initially disease-free cohort over time and applying the transition hazards. Under the important assumption made in the IPM models that there are no trends in the transition hazards, time is equal to a patient 's age. The models thus permit calculation of prevalence at a certain age from the prevalence at the previous age and the mortality and incidence in the age interval.

Although they are based on this common conceptual model, the actual model calculations differ. DisMod uses a set of linear differential equations that describe the transitions between the states. The solution of the equations is approximated by using the finite differences method. Incidence and case-fatality hazards are required as input parameters, and we approximated them using rates. Case-fatality rates were calculated from mortality data and the prevalence calculated from our own model (see below and Annex). Since DisMod cannot calculate data for age groups over 90 years and can only handle a limited number of age groups, we specified 5-year age groups from 15 years to 89 years. The calculation is performed using a competing risk life table [6]. General mortality data for the Netherlands for 1991–95 reported by CBS were specified.

Figure 2.1. **Schematic representation of a Markov model for cancers**



Our model gives an exact solution based on an analytical solution of a continuous time Markov process [7]. We refer to it here as the analytical model. Using a spreadsheet we implemented the formula for the calculation of prevalence from incidence and mortality probabilities (see Annex). Mortality and incidence rates per 5-year age group up to ≥95 years served as input. We first interpolated these data to 1-year age groups up to ≥95 years and then converted them to probabilities (see Annex for formulas and methods). Apart from the different calculation method (approximate versus exact), the analytical model thus also differs from DisMod in the way it treats mortality. Mortality probability relates to the total population, whereas case-fatality, used in DisMod, concerns only prevalent cases. In the event of inconsistent data, the mortality probability may exceed the predicted prevalence, resulting in negative prevalence estimates in the analytical model, which is not possible if case-fatality is used.

The models were assessed by comparing the calculated prevalence with the observed data. We extrapolated the observed prevalence data from ≥85 years to ≥95 years by applying the cubic-spline methodology and using life-table derived mean ages of 89.6 years and 90.8 years, for men and women aged ≥85 years, respectively.

Results

Application of the internally consistent MISCAN data to the models resulted in prevalence estimates that were virtually identical to those generated by MISCAN and to one another. When the observed data were applied the results of the two models were also practically identical. The reproducibility of MISCAN data and the consistency of the results from the two models suggest that they calculated the correct results.

The prevalences calculated from national data by the analytical model are shown with the observed prevalences in Figures 2.2–2.5. The model estimates increase with age, at first exponentially, but subsequently at a slower pace. At ages >80 years the estimates reach a maximum and then decline. The decline at older ages is most apparent for stomach cancer, the calculated prevalence decreasing to zero or even to negative values; the smallest decline is that for breast cancer.

Figure 2.2. **Observed and estimated prevalence of female breast cancer by age.** Empirical data (broken line) are from the IKZ region and were interpolated from 5- to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid line) are calculated from national incidence and mortality data using the analytical model.

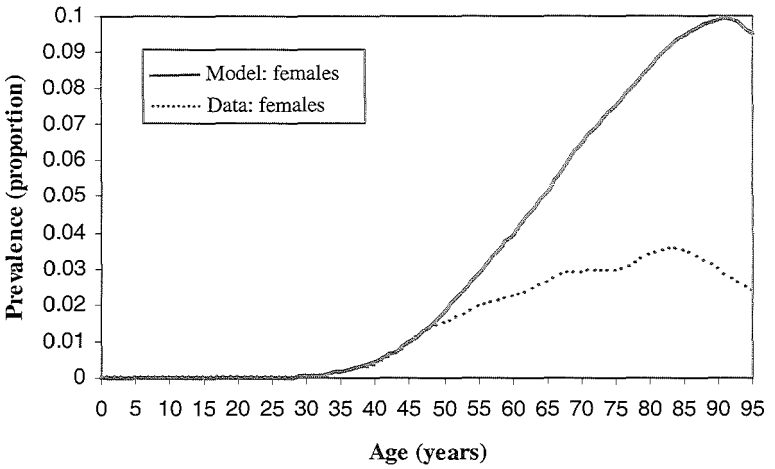


Figure 2.3. **Observed and estimated prevalence of prostate cancer by age.** Empirical data (broken line) are from the IKZ region and were interpolated from 5- to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid line) are calculated from national incidence and mortality data using the analytical model.

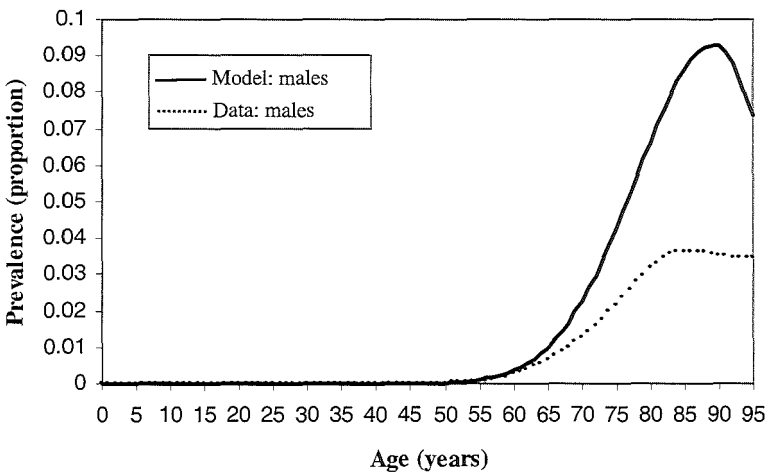


Figure 2.4. **Observed and estimated prevalence of colorectal cancer, by age and sex.** Empirical data (broken lines) are from the IKZ region and were interpolated from 5- to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid lines) are calculated from national incidence and mortality data using the analytical model.

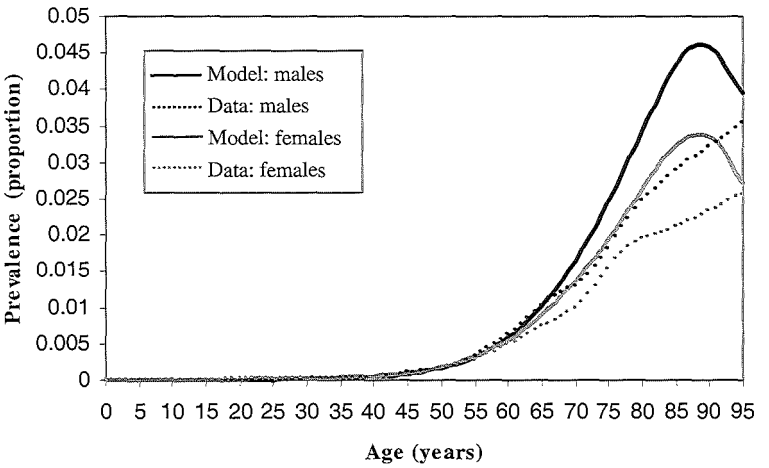
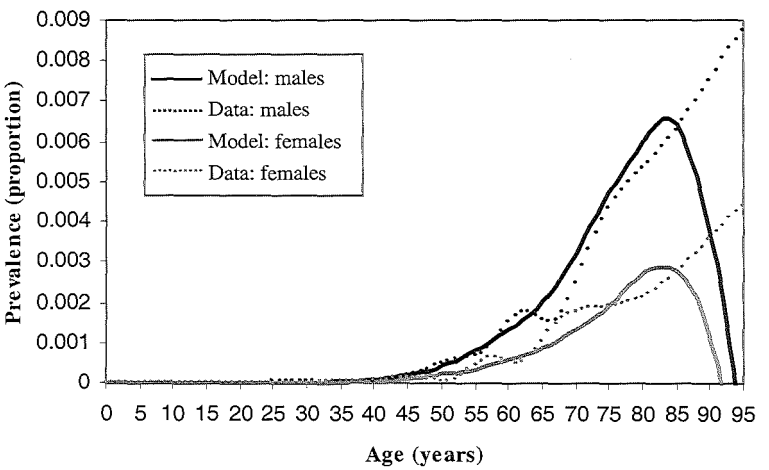


Figure 2.5. **Observed and estimated prevalence of stomach cancer, by age and sex.** Empirical data (broken lines) are from the IKZ region and were interpolated from 5- to 1-year age groups and extrapolated for ages >85 years using the cubic-spline method. Model estimates (solid lines) are calculated from national incidence and mortality data using the analytical model.



The predicted prevalences are nearly always larger than the observed ones. However, stomach cancer is exceptional in this respect in that the estimate approximates to the observed value, except for ages >85 years. For prostate and breast cancer the discrepancy is large (depending on age, the model calculations are up to about two and three times larger, respectively), while for colorectal cancer it is intermediate (up to about 1.5 times larger).

Discussion

The results from the artificial data sets support the validity of the IPM models: despite the difference of a lognormally distributed survival in MISCAN and a piecewise exponentially distributed survival in the IPM models, the latter are able to reproduce the MISCAN prevalence very well. In addition the two IPM models produce virtually the same results, a further indication of validity. Nevertheless, when registered incidence and mortality data were used, the predicted prevalence differed considerably from that observed, and for stomach cancer impossible results were produced. Three possible reasons for these discrepancies are considered below.

Regional differences

The regionally observed prevalence data, to which the national prevalence estimates are compared, may not be representative of the national situation. Breast cancer screening was introduced in the IKZ region between 1993 and 1997, whereas in the rest of the Netherlands it was introduced around 1990. Cause-specific prostate cancer mortality is unequally distributed [8]. We explored the influence of regional differences by applying regional incidence and mortality data to the model. A comparison of the results of these calculations (not shown) with the empirical data revealed that the differences between estimates and observations were similar, and, if anything, somewhat larger than the discrepancies we found on using national data in the models. The regional variation in disease epidemiology therefore could not explain the differences.

Past trends

Because both models are based on the assumption that incidence and mortality are in a steady state, the occurrence of trends in incidence or mortality would lead to discrepancies between the model estimates and the observations. Prevalence is a stock variable, comprising all past incident cases that are still

alive. It is therefore dependent on incidence and case-fatality from the past as well as the present.

Cancer incidence has a tendency to rise for the tumours we studied, because of increased awareness and screening, and for other, unknown, reasons [9, 10]. The incidence of breast cancer, for example, is presumed to show a secular trend of 1% per year [11], on top of which an additional increase is imposed because of the introduction of breast cancer screening around 1990 in the Netherlands. A notable exception to this rising incidence is presented by stomach cancer, for which there has been a long-term secular decline [9, 10].

Cancer mortality, meanwhile, remains relatively stable [9, 10, 12-14], except for stomach cancer and female colorectal cancer, for which it has declined. With increasing incidence and constant mortality, prevalence increases over time but does so less rapidly than incidence, since it also includes persons who became incident in the past. Consequently, applying current incidence and mortality to the model produces estimates that are higher than the observations. The largest deviations were seen concordantly for breast cancer and prostate cancer, the two cancers for which the rise in incidence has been most apparent, mainly because of screening.

In an additional analysis old incidence data for breast cancer were used in the model in order to check whether trends could explain the discrepancies. The risk of mortality for breast cancer patients remains elevated for more than 20 years after diagnosis. We therefore used regional incidence data for 1968–72 [12], obtaining an estimated prevalence close to the observed value for 1993. This shows that the trend in incidence may indeed cause a difference. Nevertheless it cannot explain the discrepancy entirely: the average incidence that cases prevalent on 1 January 1993 were exposed to lies somewhere between the 1991–95 and the 1968–72 incidences. Although we believe the effect of the trends in incidence is considerable, other factors evidently also play a role.

Data inaccuracies

Inaccuracies in the epidemiological estimates are the third possible reason for the differences between observations and model predictions. Statistics on mortality by cause-of-death in the Netherlands are assumed to be reliable, although no studies are known in which the completeness of the death registry has been investigated in absolute terms. Compared with other European countries, in the Netherlands the detection fraction for cancer as a cause of death

is high [15]. Furthermore, it has been argued that deaths from cancer in general are not likely to be missed [16], although misclassification between cancers may occur for older age groups. Thus it is unlikely that underregistration of cancer deaths is a causative factor in our generally higher prevalence estimates, especially with regard to young and middle-aged people. Nevertheless, the underestimation of mortality remains a possible explanation for discrepancies. Since we did not include excess mortality from other diseases in our model we implicitly assumed it to be zero. However, cancer patients also suffer from an increased risk of dying from diseases other than cancer [17]. We believe that, in addition to the effect of trends, the impact of ignoring this factor makes an important contribution to the discrepancies.

At older ages, where multiple medical problems are frequent and pathological examinations are performed relatively infrequently, misclassification of cancer deaths may lead to the overregistration of deaths for the more frequent types of cancer. This would cause prevalence estimates to be too low and could contribute to the decline of our prevalence estimates at older ages.

Cancer incidence data in the Netherlands are reliable. Nevertheless, because they are based on pathology and hospitalization data, those incident cancer cases that did not undergo a pathological examination and were not hospitalized would be systematically excluded from registration. It has been estimated that this would lead to an underregistration of 1.3–1.6% [18, 19]. Moreover, some cases that are included in pathology or hospitalization registries are missed. This non-systematic exclusion has been estimated to occur in 2.2% of cases [19]. A completeness of approximately 96.2%is thus achieved, which is comparable to the level of completeness in several other national cancer registries [19]. This incompleteness seems to be concentrated in the highest age groups; one study suggested that missed incident cases mostly related to elderly persons with cancer of the digestive tract [18], although this was not confirmed in another study [19]. The underregistration of incidence may help to explain the impossible negative prevalence calculated for stomach cancer and the decline in the estimates for colorectal cancer at older ages (Figure 2.4 and 2.5).

Underregistration cannot, however, explain the finding that the prevalence estimates are generally higher than the observations for the other age groups. Multiple malignancies can contribute to this. The incidence registry counts the number of malignancies, whereas the prevalence data are based on

the number of persons with a malignancy. Consequently, a person with multiple malignancies in the same organ is counted more than once in the incidence data but only once in the prevalence data. For breast and colorectal cancer such multiple malignancies may be present, and can account for up to 10% and 15%, respectively, of the incident cases (J.-W. Coebergh, personal communication, 2000). This would make our prevalence estimate too high and would explain part of the differences, but not more than the 10% or 15% by which the incidence is overestimated.

The incompleteness of prevalence data could also be a factor contributing to the higher estimates. Although based on regional incidence data, prevalence data may be somewhat less complete because cancer registration was less complete in its early years than more recently [19]. Since only old cases are underestimated in this way, prevalence is only affected if the survival time is long. Furthermore, this underestimation might be diminished by the opposite phenomenon: overestimation of prevalence resulting from incomplete ascertainment of survival status. The latter incompleteness would be very small, however, since deaths are unlikely to be missed, although problems may arise when persons have moved out of the country. We believe that the underestimation of prevalence data is not large and that it is unlikely to explain a large part of the differences.

Conclusion

The test with the artificial data supports the formal validity of IPM models. However, the confrontation with the four empirical data sets of presumed high quality shows that, in practice, there may be large discrepancies between measurements and calculations. The discrepancies are likely to be attributable in considerable measure to past trends in incidence but also to data inaccuracies, the most important source of which seems to be underestimation of mortality as a result of ignoring excess mortality from other causes.

The model cannot distinguish between the effects of trends and the effects of data inaccuracies. Separating these effects would require a dynamic model that describes the disease processes over time, and could incorporate the effects of past trends. Unfortunately, such a model would be much more complex. Moreover, since the trends would have to be quantified, more input

data would be required, and these have proved difficult to obtain. Consequently, a dynamic analysis is often not feasible.

In practice use of IPM models such as DisMod occurs particularly when data are incomplete and/or of low quality. In such circumstances it is impossible to distinguish between the apparent inconsistencies that represent real data problems and those that are attributable to past trends. This complicates the use of such models in improving estimates of disease epidemiology. Considerable judgement has to be exercised when the disadvantage of forcing data to comply with the assumption of a steady state is weighed against the goal of reducing the unreliability of the data. Expert knowledge on disease epidemiology and registries remains indispensable for guiding this process.

Acknowledgements

We thank Dr Jan-Willem Coebergh, Dr Marie-Louise Essink-Bot, Dr Pieter G.N. Kramers, Professor Paul J. van der Maas and Professor Daan Kromhout for their constructive suggestions and comments on this manuscript. The present work was sponsored by the Netherlands Institute of Health Sciences.

References

1. Rosenberg PS, Biggar RJ, Goedert JJ, Gail MH. Backcalculation of the number with human immunodeficiency virus infection in the United States. *Am J Epidemiol* 1991;133(3):276-85.
2. Murray CJ, Lopez AD. Quantifying disability: data, methods and results. *Bull World Health Organ* 1994;72(3):481-94.
3. Barendregt JJ, Baan CA, Bonneux L. An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus. *Epidemiology* 2000;11(3):274-279.
4. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. *Comput Methods Programs Biomed* 1985;20(1):79-93.
5. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. *Comput Biomed Res* 1999;32(1):13-33.
6. Manton KG, Stallard E. *Chronic disease modelling*. 1 ed. New York: Oxford University Press; 1988.

for age n by averaging the results of two successive ages. This can be given in either 1-year or 5-year age groups.

The DisMod model

The DisMod model can be downloaded from the Internet (at URL: http:// www.hsph.harvard.edu/organizations/bdu/dismod/index.html). (Two versions of DisMod are available at this URL –we used DisMod I in our analyses.) For this model the same assumptions of steady state, constant hazards in the age interval and independence of other-cause mortality are required.

We approximated input hazards by rates. The appropriate input formats were calculated using the following expressions:

$$CFR_n = -LN\left[1-\left\{\left(m_n - r_n * \left(1-p_n\right)\right)/\, p_n\right\}\right], \text{ if } m_n<p_n, \text{ otherwise } 0 \qquad \text{eq.}\{3\}$$

$$r_n = \left[IR_n * \left(1-\exp\left(-CFR_{n-1}\right)\right) - CFR_{n-1} * \left(1-\exp\left(-IR_n\right)\right)\right]/\left[IR_n - CFR_{n-1}\right],$$

where CFR_n is the case-fatality rate for age group n and r_n is the probability of making two transitions in age group n (the other parameters are mentioned above).

Conversions to the appropriate input formats were made by using the interpolated data (1-year age intervals). We used the prevalence calculated by the previous model to compute case-fatality. Because the formula for calculating r_n requires the CFR_n as input data, and vice versa, we used the case-fatality rate of the previous age (CFR_{n-1}) in this calculation. This produced a slight deviation.

We then back-transformed the input data to 5-year age groups. DisMod was then used to calculate annual incidence and mortality rates and mean prevalence per 5-year age group.

3

Estimating the prevalence of breast cancer using a disease model: data problems and trends

Michelle E. Kruijshaar, Jan J. Barendregt, Lonneke V. van de Poll-Franse. Estimating the prevalence of breast cancer using a disease model: data problems and trends. *Population health metrics* 2003; 1 (1): 5.

Abstract

Background:

Health policy and planning depend on quantitative data of disease epidemiology. However, empirical data are often incomplete or are of questionable validity. Disease models describing the relationship between incidence, prevalence and mortality are used to detect data problems or supplement missing data. Because time trends in the data affect their outcome, we compared the extent to which trends and known data problems affected model outcome for breast cancer.

Methods:

We calculated breast cancer prevalence from Dutch incidence and mortality data (the Netherlands Cancer Registry and Statistics Netherlands) and compared this to regionally available prevalence data (Eindhoven Cancer Registry, IKZ). Subsequently, we recalculated the model adjusting for 1) limitations of the prevalence data, 2) a trend in incidence, 3) secondary primaries, and 4) excess mortality due to non-breast cancer deaths.

Results:

There was a large discrepancy between calculated and IKZ prevalence, which could be explained for 60% by the limitations of the prevalence data plus the trend in incidence. Secondary primaries and excess mortality had relatively small effects only (explaining 17% and 6%, respectively), leaving a smaller part of the difference unexplained.

Conclusion:

IPM models can be useful both for checking data inconsistencies and for supplementing incomplete data, but their results should be interpreted with caution. Unknown data problems and trends may affect the outcome and in the absence of additional data, expert opinion is the only available judge.

Background

Estimates of disease-specific incidence, prevalence and mortality, specified by age and sex, are important information to health care policy and planning. They are essential inputs to cost-effectiveness analyses and burden of disease calculations. Empirical data, however, are often difficult to obtain or are of questionable validity. To remedy some of these data problems, disease models have been developed that describe the relationship between the epidemiological parameters, by exploiting the causal structure of a disease. Incidence, prevalence, mortality models (IPM models) formalise the relationship between the three parameters, using the fact that incidence has to precede prevalence, and that cause-specific mortality can only follow disease. IPM models have been used frequently both to supplement missing data and to study the agreement between different epidemiological data [1-4]. Our previous study supported the formal validity of IPM models, but when the modelling was applied to empirical data on four types of cancer, the model calculations differed to a large extent from the empirical data [5]. For breast cancer the difference was particularly large. It was argued that these discrepancies may indicate inconsistencies in the data, but that they may also be caused by time trends.

When the data for one disease are not in accordance with each other, they are internally inconsistent. Inconsistencies may be caused by differences in the completeness of the data. For example, when more incident cases are missed than deaths, incidence and mortality are inconsistent. Also, inconsistencies may arise when the data were derived from different contexts (e.g. a different region) or measured differently (e.g. varying case-definitions). Applying inconsistent incidence and mortality to an IPM model will result in under- or overestimating prevalence, and thus in discrepancies between model estimations and empirical prevalence data. Time trends, on the other hand, may cause the data to appear inconsistent in a steady state model, while in fact they are not. Because prevalence is the resultant of incident cases from the past, it cannot react instantaneously to changes in incidence and case-fatality, but only with a certain delay. It is possible to account for the effects of time trends in a dynamic model, but this requires additional input data on the nature and size of the trends, which are not available for most diseases. Often, we do not even know whether a trend is present or not, and the researcher faces a dilemma what to do with the

discrepancies. Adjusting observed data for apparent inconsistencies that are in fact the consequence of past trends would rather defeat the purpose of IPM models.

For breast cancer in the Netherlands the discrepancy between observed prevalence and prevalence calculated from incidence and mortality was particularly large [5]. Fortunately, for breast cancer several data problems are known, and, in addition, there is a tentative estimate of the trend in incidence. This allowed us to quantify the relative contribution of the trend and several known data problems on the discrepancy and to throw some light upon the researcher's dilemma. Even though prevalence of breast cancer is not a very useful epidemiological measure, it does allow us to illustrate the difficulties in the use of IPM modelling because of the relative abundance of data.

Methods

General approach

We calculated the point-prevalence of breast cancer in the Netherlands and its 95% confidence interval from national incidence rates and cause-specific mortality rates using the IPM model described by Barendregt et al. [6]. We compared it to regional prevalence data. To determine the separate effects of trends and known data problems we next recalculated the model:

1) incorporating the incompleteness in the prevalence data (see below),

2) adjusting for a trend in incidence,

3) adjusting for double counting of incident women with secondary primaries, and

4) taking excess mortality from non-breast cancer deaths into account.

For point c a single cohort model was used, while for a, b, and d it was necessary to use a multi-cohort model that takes into account both age and calendar time. We estimated the proportion of the difference that was explained by each of the recalculations from the overall differences in the number of prevalent cases using 1993 population figures and summing over all age groups.

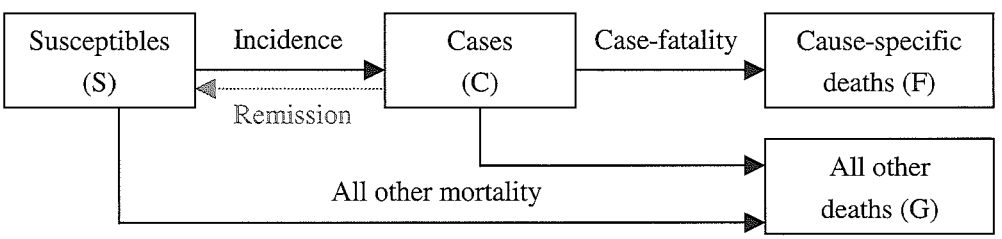
IPM model

Our model is based upon the conceptual disease model depicted in Figure 3.1 and was described in Kruijshaar et al. [5]. Briefly, the model describes a population as being in two states: diseased or susceptible, while transition hazards determine how people move from one state to another. Following an initially disease-free cohort over time and applying the transition hazards, the number of cases can be calculated. Under the important assumption of a steady-state situation, time is equivalent to age. The model then allows the calculation of prevalence at a certain age from the prevalence at the previous age and the transition hazards. By assuming furthermore equal mortality from other causes in cases and susceptibles, prevalence at exact age n can be calculated from incidence and cause-specific mortality probabilities using formula 1a in the Appendix [6].

Baseline calculation

First, we calculated prevalence from national incidence and mortality rates of female breast cancer (ICD-9 code 174) for 1991-1995, averaged by five-year age groups up to 85+. Incidence data obtained from the Netherlands Cancer Registry (NKR) are based on pathology and hospital admission data, and the mortality data from Statistics Netherlands (CBS) on death certificates. To enable comparison with the prevalence data we used incidence rates excluding in-situ tumours. Incidence and mortality rates were first interpolated to one-year age groups using the cubic-spline method and then converted into probabilities (see the Appendix for formulas). A 95% confidence interval of the calculated

Figure 3.1. **A Markov model for cancers**



prevalence was obtained by parametric bootstrapping assuming a Poisson distribution for numbers of incident cases and deaths in each age group. The @risk software programme [7] was used to simulate 10000 iterations by Monte Carlo sampling.

The Regional Cancer centre South (IKZ) was the only cancer registry in the Netherlands that has estimated prevalence. The IKZ determined prevalence for a specific part (the core region) of the region South at 1-1-1993 by checking the vital status of all incident cases registered in the region since 1970 against the municipal population administration and the National Death Index. If a person had moved to another municipality in the Netherlands, data from that municipal population administration was used (under Dutch law, registration with the municipal population administration is obligatory within five days of changing address). In case of migration to another country, a case was lost to follow up. In-situ tumours are not included. For more information on these data we refer to Coebergh et al. [8]. A 95% confidence interval around these data (by five-year age group up to 85+) was calculated assuming a binomial distribution (see formula 2 in the Appendix). Point estimates and confidence intervals were interpolated to one-year age groups as described in the Appendix.

Limitations of the prevalence data

Prevalence registered by the IKZ does not include patients diagnosed before 1970 and is consequently underestimated. To quantify the effect of this, we created a dynamic model, incorporating a parameter y that represents the number of years prior to incidence. Prevalence in the reference year $y0$, was calculated from prevalence in the 23 years prior to $y0$ (see Appendix).

Also, prevalence data refer to a specific region only, whereas the model calculations are based on national input. We showed in our previous study [5], that differences between regional and national incidence and mortality rates hardly affected the calculated prevalence. Therefore, we did not further examine this here.

Trend in incidence

Next, we estimated the effect of a trend of increasing incidence on the calculated prevalence. While incidence increased, the population mortality rate for breast cancer has remained approximately the same in the Netherlands [8-12], thus case-fatality must have declined. Coebergh et al. have estimated the yearly rise in incidence in the region South between 1975 and 1986, to be approximately

one percent [13]. The effects of screening are not included in this estimate, as it was not introduced until later, but effects of other types of increased case finding were included –if present.

We estimated the effect using the dynamic model described above, decreasing incidence by one percent for each additional year *y* prior to the year of reference (see Appendix). We assumed the trend was present up to 95 years before the year of reference. Sensitivity to the trend was inspected, by applying a 50% higher and lower trend (1.5% and 0.5%) as input values.

Double counting incident women with a secondary primary

The NKR registers the number of incident tumours, whereas prevalence refers to women. Women are thus counted twice, if they have a second primary tumour in their breast (SP). The percentage of SPs by age group were provided for 1991-1995 by the IKZ. We estimated the incidence of women with breast cancer, by subtracting the age-specific proportion SP from the reported incidence rate and recalculated our model.

Excess mortality from non-breast cancer deaths

Breast cancer patients have been shown to be at increased risk of dying from other causes of death [14], although with longer survival (>20 years) they may experience a decreased risk [15]. Cause-specific mortality data thus underestimate the total excess mortality from breast cancer. We adjusted for this estimating excess mortality from duration-specific relative survival data assuming a lognormal survival distribution from breast cancer and a proportion not dying from breast cancer (proportion cured) [16], as described in the Appendix (formula 5). Relative survival data, i.e. excess mortality over and above the background mortality, were reported by the IKZ registry for three cohorts (1970-1979, 1980-1986 and 1987-1992) for ages < 70, and > 70 [8]. Because survival improves over time, we used survival probabilities at one, three and five years after incidence of the youngest 1987-1992 cohort. As longer follow-up was not available for this cohort, survival at 10 and 20 years was estimated using the conditional 10 and 20-year survival in the older cohorts. Fitting the lognormal model allowed us to estimate the cumulative probability of excess mortality with time after incidence, from which we calculated the yearly mortality probability (see Appendix).

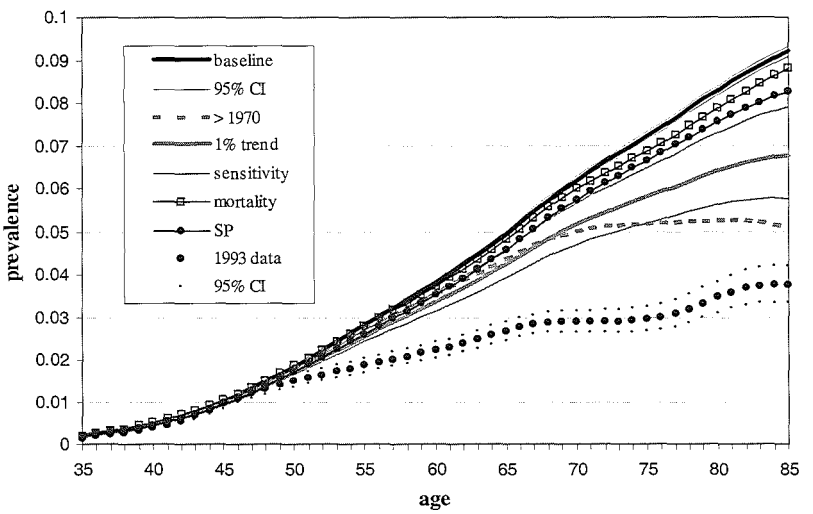
We extended our baseline model to include duration (see Appendix for the mathematical description): prevalence at age *n*, *d* years after incidence was

calculated from the prevalence at the previous age and year. Summation across all years *d* provided age-specific prevalence.

Results

In our baseline calculation we estimated the prevalence of breast cancer from national incidence and cause-specific mortality data. The results of this are shown in black in Figure 3.2 (solid line), together with the regional data (dotted line), while the total number of calculated prevalent cases and the difference with the IKZ are shown in Table 3.1. Both the IKZ and calculated prevalence increased exponentially with age to about age 47, but thereafter they started to diverge. The model calculations increased linearly to a prevalence of 9.2% at

Figure 3.2. **Calculated prevalence and prevalence data of breast cancer by age. Baseline calculation and separate effects of known data problems and trend.**



Baseline: baseline calculation, 95 % CI: 95 % confidence interval, >1970: excluding incident cases of before 1970, 1% trend: estimating the effect of a secular trend in incidence, sensitivity: 50% sensitivity borders around the effect of the secular trend, SP: adjusting for secondary primaries, mortality: including excess mortality from non-breast cancer deaths.

age 85, while the IKZ prevalence increased more slowly and levelled off to 3.7%. At age 55 the calculated prevalence was 1.5 times higher than the data, increasing to 2.5 times at age 85, a difference in total number of cases of 86%. The 95 % confidence interval for the calculated prevalence was narrow, due to the high numbers of breast cancer incidences and deaths. The uncertainty in the prevalence data was larger. From age 48 upward the confidence intervals did not overlap.

Figure 3.2 and Table 3.1 also show the effects of the known data problems and trend in incidence (coloured lines). First, excluding cases that became incident more than 23 years before the year of reference had the largest effect and altered the age pattern. Prevalence increased with age more slowly, levelling off to a maximum of 4.4% at age 78, and declining thereafter. The total number of prevalent cases was 53% higher than the IKZ prevalence, explaining 39% of the difference between calculated and IKZ prevalence. Second, incorporating a secular trend in incidence in the calculation of prevalence resulted in a lesser increase with age, levelling off after age 70, without reaching

Table 3.1. **Comparison of the total number of prevalent cases of breast cancer, estimated in different ways, and 1993 prevalence data.**

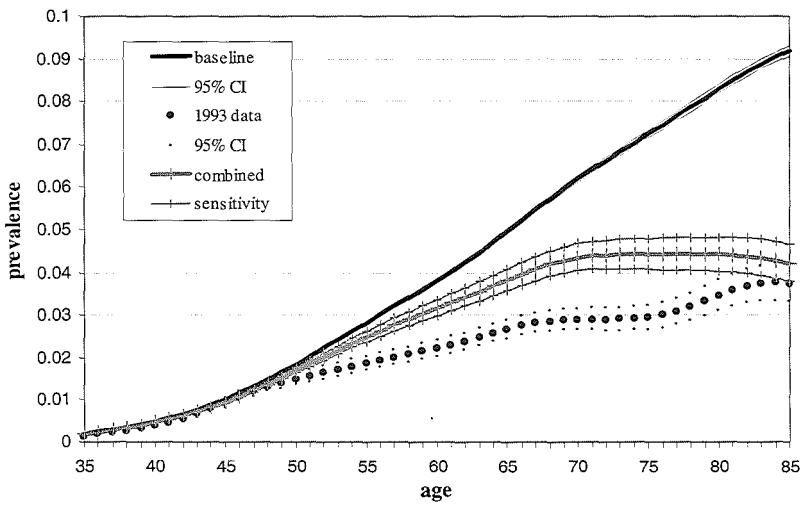
	Total number of cases estimated	% difference from 1993 data	% of the gap explained
1993 data	123216	0.0	0.0
Baseline	66370	85.6	100.0
Combined	89169	34.4	59.9
>1970	101268	52.6	38.6
1% trend	103943	56.6	33.9
SP	113440	70.9	17.2
Mortality	119769	80.5	6.1

Baseline: baseline calculation, combined: adjusting for one-% trend and excluding incident cases of before 1970, >1970: excluding incident cases of before 1970, 1% trend: estimating the effect of a secular trend in incidence, SP: adjusting for secondary primaries, mortality: including excess mortality from non-breast cancer deaths.

a plateau. The secular trend of one percent decreased the discrepancy with 44% (ranging between 49% and 18% for 50% higher and lower estimates of the trend). Third, adjusting for double counting of incident women with a second primary tumour of the breast (SP) had a smaller effect. The percentage of SP did not exceed nine percent for any age group. Subtracting this percentage from the incidence rate explained 17% of the difference. Fourth, taking excess mortality from non-breast cancer deaths into account made a difference of six percent.

The combined effect of restricting the duration of prevalence and the trends, shown in Figure 3.3, resulted in a 60% decrease of the difference (ranging between 70% and 50% for a 50% higher or lower trend). At age 85 the calculations touched upon the upper confidence limit of the data.

Figure 3.3. **Calculated prevalence and prevalence data of breast cancer by age. Baseline calculation and the effect of both the secular trend and restricting prevalence.**



Baseline: baseline calculation, 95 % CI: 95 % confidence interval, combined: adjusting for one-% trend and excluding incident cases of before 1970, sensitivity: the effect of 50% change in the estimated trend in incidence on the combined adjustment.

Discussion

We quantified the separate effects of known data problems and a trend in incidence on IPM model calculations of breast cancer prevalence, and inspected their influence on the discrepancy between model estimations and prevalence data. Two factors had a major effect on the estimated prevalence: the limitation of the IKZ prevalence data including only incident cases since 1970, and the trend in incidence. Together, they accounted for a major part (60%) of the discrepancy. Still, their combined effect leaves a part of the discrepancy unexplained. The effects of adjusting for double counting of incident women with secondary primaries of the breast and of taking excess mortality from non-breast cancer deaths into account were small. Consequently, the combined effect of all four factors does not explain the entire difference. Either unknown data problems are present contributing to the remaining difference, or we underestimated some of the effects.

The two factors with the largest effects could influence the calculations because prevalence is a stock variable: it contains cases that became incident in the past. The long survival time of breast cancer explains why these two factors can exert such a strong influence. The time lag after which prevalence fully reflects a change in incidence is determined by the rate at which the pool of prevalent cases is replaced by new cases, which, in turn, depends on the survival time. The underestimation of the prevalence data due to its limitation to incident cases after 1970 may be as large as 44% at age 85, but is lower on average, as the effect increases with age. When the effect is calculated additional to adjusting for the trend in incidence, it is also smaller: 24% at age 85. In a small part of the IKZ region prevalence can be based upon incidence registration since 1958 [9]. These data are indeed higher than the IKZ data for the larger region indicating the data will improve with longer follow-up in the future.

The estimated trend in incidence and the effect of adjusting for it are subject to some uncertainty for several reasons. First, the one percent trend is an average figure; its magnitude differs with age and calendar year. Second, the regional estimate may deviate somewhat from the national trend. Furthermore, it is a rounded figure. Nevertheless, as the results of the sensitivity analysis showed that increasing or decreasing the trend by 50% altered the percentage explained by a combined effect (of limiting prevalence and the secular trend) by only 10%-points, we expect these three effects to be small. The estimated one

percent may also be too low because it does not include the effect of the introduction of screening around 1990 in the Netherlands. We estimated this effect in an additional analysis, increasing the estimated one to five percent for the last four years. The additional effect of screening was very small. Finally, we may have overestimated the effect assuming that a one percent trend was present many years before the period for which it was estimated (1975 to 1986). Assuming no trend before 1975 in an extra analysis increased the calculated prevalence by almost 10% at age 85, but inspection of the incidence rates since 1958 for the smaller part of the IKZ region showed that incidence has increased since 1958 [9]. The effect of assuming no trend before 1960 was negligible.

The effects of adjusting for secondary primaries and excess mortality were only small, but this is not surprising. The proportion of women with a secondary primary in the breast did not exceed nine percent, preventing a much larger change in the calculated prevalence. Also, the relative risk of breast cancer patients to die from other causes of death is not that high, and may even inverse with longer follow-up [14, 15].

Thus, although the effect of the trend may be somewhat uncertain, other unknown data problems are likely to cause the remaining difference between calculated and IKZ prevalence. One possible explanation for the remaining deviation could be that when the cancer registries started completeness of incidence was not as high as it is now. As a result the prevalence is underestimated by the IKZ. On the other hand, IKZ prevalence was determined by checking population administrations, which involves matching of the registrations, which is never 100% accurate. Therefore, some deaths may have been missed and individuals inaccurately assumed to be alive, resulting in an overestimation of the IKZ prevalence. How large both counteracting effects are, is difficult to determine. An additional explanation may be found in variation between national and regional all-cause mortality, as the equations in our model are based on the assumption of similar background mortality. A higher background mortality in the IKZ region would decrease the discrepancy, but we expect only a minimal effect as the differences in all-cause mortality will be minor. Additional reasons for overestimating incidence or underestimating mortality (both resulting in overestimating prevalence) are difficult to think of. Cancer incidence data in the Netherlands are reliable with a completeness of 96.2 % around 1990 [17], and probably even higher in the years thereafter. Furthermore, mortality by cause-of-death statistics of the Netherlands are

assumed to be reliable and, compared to other European countries, the detection fraction for deaths from cancer is high [18]. The causes of the remaining discrepancy thus remain uncertain.

Conclusion

Even when data are regarded as relatively reliable, as was the case for breast cancer in the Netherlands, data problems may be present. 1993 prevalence data for breast cancer in the IKZ region are underestimated, as they do not include incident cases before 1970. Our analyses show the importance of using IPM modelling to detect data problems. However, we also showed the trend in incidence to have a large effect on the model estimations for breast cancer, complicating the use of IPM models.

IPM models can be useful both for checking for data inconsistencies and for supplementing incomplete data, but in both cases there remains the need for careful interpretation of the results. In the all too common situation where, unlike for breast cancer, no data on the size and nature of trends are available, the effects of trends cannot be estimated. Furthermore, unknown data problems may affect the model estimations in unknown directions. In the absence of additional data the researcher is faced with the dilemma of how to interpret model discrepancies, and expert opinion is the only available judge.

Acknowledgements

We would like to thank Dr. Jan-Willem Coebergh for his constructive suggestions and comments on this manuscript. This study was sponsored by the Netherlands Institute of Health Sciences.

References

1. Murray CJ, Lopez AD. Quantifying disability: data, methods and results. *Bull World Health Organ* 1994;72(3):481-94.
2. Murray CJL, Lopez AD, editors. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge, MA: Harvard School of Public Health; 1996.

3. Barendregt JJ, Baan CA, Bonneux L. An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus. *Epidemiology* 2000;11(3):274-279.
4. Mathers CD, Vos ET, Stevenson CE, Begg SJ. The Australian Burden of Disease Study: measuring the loss of health from diseases, injuries and risk factors. *Med J Aust* 2000;172(12):592-6.
5. Kruijshaar ME, Barendregt JJ, Hoeymans N. The use of models in the estimation of disease epidemiology. *Bull World Health Organ* 2002;80(8):622-628.
6. Barendregt JJ, Oortmarssen GJ, van, Hout BA, van, Bosch JM, van den, Bonneux L. Coping with multiple morbidity in a life table. *Mathematical Population Studies* 1998;7(1):29-49.
7. Palisade C. *@RISK, advanced risk analysis for spreadsheets*. 4th ed. Newfield, NY, USA: Corporation Palisade; 2000.
8. Coebergh JWW, van der Heijden LH, Janssen-Heijnen MLG. *Cancer Incidence and Survival in the Southeast of the Netherlands 1955 - 1994: a report from the Eindhoven Cancer Registry.* Eindhoven: Integraal Kankercentrum Zuid; 1995.
9. Coebergh JWW, Janssen-Heijnen MLG, Louwman WJ, Voogd AC. *Cancer. Incidence, care and survival in the Southeast of the Netherlands 1955 - 1999: a report from the Eindhoven Cancer Registry (IKZ) with cross-border implications.* Eindhoven: Integraal Kankercentrum Zuid; 2001.
10. Maas IAM, Gijsen R, Lobbezo IE, Poos MJJC. *[Public health status and forecast 1997. Part 1. The health status: an update].* Bilthoven: Rijksinstituut voor Volksgezondheid en Milieu; 1997 (in Dutch).
11. Visser O, Coebergh JWW, Schouten LJ, van Dijck JAAM. *Incidence of Cancer in the Netherlands 1995.* Utrecht: Vereniging van Integrale Kankercentra; 1998.
12. Visser O, Coebergh JWW, Schouten LJ, van Dijck JAAM. *incidence of cancer in the Netherlands 1997.* Utrecht: Vereniging van Integrale Kankercentra; 2001.
13. Coebergh JW, Crommelin MA, Kluck HM, van Beek M, van der Horst F, Verhagen-Teulings MT. [Breast cancer in southeast North Brabant and in North Limburg; trends in incidence and earlier diagnosis in an unscreened female population, 1975-1986]. *Ned Tijdschr Geneeskd* 1990;134(15):760-5 (in Dutch).
14. Brown BW, Brauner C, Minnotte MC. Noncancer deaths in white adult cancer patients. *J Natl Cancer Inst* 1993;85(12):979-87.
15. Louwman WJ, Klokman WJ, Coebergh JW. Excess mortality from breast cancer 20 years after diagnosis when life expectancy is normal. *Br J Cancer* 2001;84(5):700-3.

16. Rutqvist LE. On the utility of the lognormal model for analysis of breast cancer survival in Sweden 1961-1973. *Br J Cancer* 1985;52(6):875-83.
17. Schouten LJ, Hoppener P, van den Brandt PA, Knottnerus JA, Jager JJ. Completeness of cancer registration in Limburg, The Netherlands. *Int J Epidemiol* 1993;22(3):369-76.
18. Mackenbach JP, Van Duyne WM, Kelson MC. Certification and coding of two underlying causes of death in The Netherlands and other countries of the European Community. *J Epidemiol Community Health* 1987;41(2):156-60.
19. C Radhakrrishna Rao: Linear statistical inference and its applications, 2nd edn. New York: Wiley; 1973.

APPENDIX

The steady state IPM model

Assuming a steady state situation and equal other-cause mortality for cases and non-cases, prevalence at exact age n, can be calculated from incidence and cause-specific mortality data, using [6]:

$$p_n = [p_{n-1} - m_{n-1} + i_{n-1} * (1 - p_{n-1})]/(1 - m_{n-1}), \quad (1a)$$

where m_{n-1} is the mortality probability of age group *n-1*; and i_n is incidence probability of susceptibles of that age group. The formula gives an exact solution of a continuous time Markov process.

The input parameters for this formula are mortality probability at age group n and the incidence probability of susceptibles at that age group. We interpolated incidence and mortality rates from five-year to one-year age groups using the cubic-spline method adopting a mean age of 90.8 for women aged 85 and over that was derived from a life table corresponding to general mortality. We assumed these rates approximate the hazards fairly well. Hazards were then converted to probabilities:

$$probability_n = 1 - EXP(-hazard_n). \quad (1b)$$

The incidence rate among susceptibles was calculated from national incidence hazards using the prevalence of the previous age group, resulting in a slight deviation:

$$IR_{n,susceptibles} = IR_{n,national} / (1 - p_{n-1}), \quad (1c)$$

where $IR_{n,susceptible}$ is the incidence rate among susceptibles, and $IR_{n,national}$ the incidence rate within the national population.

Confidence interval for prevalence data

Assuming prevalence data are binomially distributed, their 95 % confidence interval can be calculated from [19]:

$$CI95_n = \sin\left[\arcsin\sqrt{p_n} \pm 1.96\sqrt{1/(4N_n)}\right]^2, \quad (2)$$

with N_n representing the total number of women in age-group n: $N_n = C_n + S_n$.

A dynamic IPM model

The baseline IPM model is extended to incorporate the parameter y, the number of years prior to the reference year:

$$p_{n,y} = \left[prev_{n-1,y+1} - mort_{n-1} + inci_{n-1} * \left(1 - prev_{n-1,y+1}\right)\right] / \left(1 - mort_{n-1}\right). \quad (3)$$

To correct for the fact that more than 23 years before the year of reference incidence was not measured, incidence, and thus prevalence, was set to zero for y>23. To correct for a one-% trend in incidence, incidence was decreased by one-% for each additional year prior to the reference year (maximum 95):

$$p_{n,y} = \left[prev_{n-1,y+1} - mort_{n-1} + inci_{n-1} * 0.99\text{^}\left(y+1\right) * \left(1 - prev_{n-1,y+1}\right)\right] / \left(1 - mort_{n-1}\right). \quad (4)$$

Describing survival from breast cancer as lognormally distributed with a proportion cured

We estimated excess mortality with breast cancer from duration-specific relative survival data assuming a lognormal survival with a proportion cured [16].
When $RSurv_{n,d}$ is the relative survival in age-group n, d years after incidence, c_n is the proportion cured for that age-group, and μ_n and σ_n are the parameters of the cumulative lognormal distribution (*Logndist*), the model can be described as:

$$RSurv_{n,d} \approx \left(1 - c_n\right) * \left(1 - Logndist\left[\mu_n, \sigma_n, d\right]\right) + c_n . \quad (5a)$$

Fitting this model to the survival data allows the estimation of the cumulative mortality at different years d after incidence. Non-cumulative mortality at exact d years after incidence is calculated by:

$$mortyr_{n,d} = \left(1 - c_n\right) * \left(Logndist\left[\mu_n, \sigma_n, d+1\right] - Logndist\left[\mu_n, \sigma_n, d\right]\right). \quad (5b)$$

A duration specific IPM model

Prevalence at age n, d years after incidence can be calculated from the prevalence at the previous age and year, using:

$$p_{n,o} = i_n , \qquad \text{for d=0, and} \quad (6a)$$

$$p_{n+1,d+1} = p_{n,d} * \left(1 - mortyr_{n,d} * \left(1 - c_n\right)\right) / \left(p_{n,d} - mortyr_{n,d} * \left(1 - c_n\right)\right), \text{ for d>0.}$$

Summation across all years d results in an estimate of age-specific prevalence:

$$p_n = \sum_{d=0}^{\infty} p_{n,d} . \quad (6b)$$

4

Lifetime prevalence estimates of major depression: an indirect estimation method and a quantification of recall bias

Michelle E. Kruijshaar, Jan J. Barendregt, Theo Vos, Ron de Graaf , Jan Spijker, Gavin Andrews. Lifetime prevalence estimates of major depression: an indirect estimation method and a quantification of recall bias. *Submitted.*

Abstract

Objective

Measurement of lifetime prevalence of depression in cross-sectional surveys is biased by recall problems. We estimated it indirectly for two countries using modelling and quantified the underestimation in the empirical estimate for one.

Method

We used a microsimulation model to generate population-based epidemiological measures of depression. We fitted the model to 1- and 12-month prevalence data from the Netherlands Mental Health Survey and Incidence Study (NEMESIS) and the Australian Adult Mental Health and Wellbeing Survey.

Results

The lowest proportion of cases ever having an episode in their life is 30% of men and 40% of women, for both countries. This corresponds to a lifetime prevalence of 20% and 30%, respectively, in a cross-sectional setting (aged 15-65). The NEMESIS data were 38% lower than these estimates.

Conclusion

Modelling enabled us to estimate lifetime prevalence of depression indirectly. This method is useful in the absence of direct measurement, but also showed that direct estimates are underestimated by recall bias and by the cross-sectional setting.

Introduction

Major Depression (MD) is a debilitating [1-6] and prevalent disease, which is among the top five leading causes of burden of disease worldwide [7, 8]. In recent community surveys, four to ten percent of the general population was shown to experience an episode of MD within a year [9-13]. Another way to express the prevalence of MD is lifetime prevalence, which is defined as the proportion of people ever having experienced at least one episode. Ideally this would be measured based on completed life courses, but since community studies have to rely on self-report, this is not possible. In practice, studies measure a cross-sectional life time prevalence, ignoring incident cases that appear after the survey takes place. Two recent surveys reported lifetime prevalences of over 15% [10, 11].

These recent surveys rely on structured diagnostic interviews that use the respondents' reporting of depressive symptoms to diagnose MD. To collect lifetime prevalence data, respondents have to recall the presence and co-occurrence of symptoms retrospectively over their past lifetime, possibly many years after they occurred. In such an assessment, problems with recall are not uncommon [14-19]. A comparison of cross-sectional and longitudinal data suggested that lifetime prevalence based on recall may be severely underestimated [20]. Furthermore, one study found that 25 years after admission for MD, 50% of the patients were not detected by the Composite International Diagnostic Interview (CIDI), the interview used in the recent surveys [21]. This recall problem was an important reasons to refrain from measuring lifetime prevalence in the Australian Mental Health and Wellbeing Survey (personal communication).

Instead of collecting lifetime prevalence empirically, it can also be estimated indirectly, using a modelling approach. It can be reconstructed from survey data, which are less prone to recall bias, such as 1- and 12-month prevalence. This approach enables us to check the extent to which existing measures of lifetime prevalence are underestimated by recall bias, but also provides an estimate in the absence of data.

In this study we indirectly estimate the lifetime prevalence of MD on data from the Australian Adult Mental Health and Wellbeing Survey [9] and the Netherlands Mental Health Survey and Incidence Study (NEMESIS) [10]. In

addition to providing an estimate for Australia, where empirical data is missing, a comparison of the indirect estimate with the empirical NEMESIS data provides a quantification of recall bias.

Methods

The data

The three-wave longitudinal NEMESIS survey was based on a random sample drawn from the Dutch general adult population, aged 18 to 64 [10, 22]. Using the CIDI [23, 24] version 1.1, diagnoses were derived according to third revised version of the Diagnostic and Statistical Manual of Psychiatric Disorders (DSM-III-R) [25]. We used 1- and 12-month prevalence data from wave I (1996). Total incidence (of both first and recurrent cases) in the 12 months between wave I and II were used to check the model outcome. For consistency with the incidence data, prevalence data were derived from the sample that participated in both wave I and II (5018 people) and exclusion criteria were not applied.

The cross-sectional Australian Mental Health and Wellbeing survey was based on a household sample of adults 18 years and older. Using the CIDI version 2.1, 1-month and 12-month prevalence was assessed for 1997. Both DSM-IV and ICD-10 diagnoses were available. To be consistent with the Dutch data, we used the DSM diagnoses, without application of exclusion criteria, and excluding subjects older than 65.

The weighted data of the surveys were used. For more information on the methods, sampling and response we refer to Bijl et al. [22] and Andrews et al. [9]. For both studies 95% confidence intervals around the prevalence data were calculated assuming a binomial distribution, using:

$CI95_a = \sin\left[\arcsin\sqrt{p_a} \pm 1.96\sqrt{1/(4N_a)}\right]^{\wedge 2}$, where p_a is the prevalence at age-group a and N_a the total number of persons at that age-group [26].

The model

To allow for the large degree of heterogeneity that is seen in MD, we used a microsimulation approach. This technique describes the disease process of an individual in terms of probabilities and their distributions [27]. Individual life histories are generated by random drawings from these distributions. Adding a large number of life histories creates a population from which population based epidemiological measures, such as lifetime prevalence and number of episodes,

can be derived.

We defined distributions for all the relevant transitions. The following algorithm creates a life history for each person in the population:

1. First an age at death is drawn from the observed survival curve (Statistics Netherlands and Australian Bureau of Statistics).
2. A second random draw from a uniform distribution determines whether the person is susceptible for depression or not (parameter "susceptible"). If so we continue with point 3, else all years are lived free of MD.
3. From a Weibull distribution, a standard distribution to model 'time to failure', age at first incidence is drawn (parameters: $\alpha 1$ and $\beta 1$). If this age is lower than the age at death, we continue with point 4, else all years are lived free of MD.
4. We assumed the duration of a depressive episode to be lognormally distributed [28] (parameters: μ and σ). Random draws are made until the episode lasts at least two weeks. If age at death is higher than age at incidence increased with duration, we continue with point 5, else the duration of the episode is reduced to last until death. If it becomes shorter than two weeks, it is dropped.
5. From a second Weibull time until a repeat episode is drawn (parameters: $\alpha 2$ and $\beta 2$). Eight weeks, by definition the a minimum period between episodes is added to the randomly drawn time to the next episode. If the age of incidence is lower than age at death, we continue with point 6, else all remaining years are lived free of MD.
6. Points 4 and 5 are repeated until the age at death is reached.

The model is thus defined by seven parameters.

Fitting the model to survey data

First we fitted the parameters of the lognormal distribution to sex-specific mean duration of episodes (u) calculated from 12- and 1-month prevalence from: $u = \frac{365 * p_w}{y * n} + 2i - w + 1$, where u is the mean duration of episodes in days; p_w the 1-month prevalence, y the 12-month prevalence, i the minimum length of episode minus 1 day (13 days); w the length of the observation window (30.4 days for 1-month prevalence); and n the mean number of incident episodes of average duration per year for people being 12-month prevalent. We assumed n is one. Next, we fitted the remaining five parameters to the prevalence data

specified by 5-year age group and sex. The fitting procedures used Nelder and Mead's downhill simplex method [29].

The available data do not provide sufficient restrictions for only one optimal fit. Various combinations of lifetime prevalence and lifetime number of episodes lead to the same 1- and 12-month prevalence, although the number of combinations is restricted by the age pattern of the prevalence data. We choose those fits with the lowest lifetime prevalence in order to estimate the minimum extent of the underestimation.

A cross-sectional estimate and excess mortality

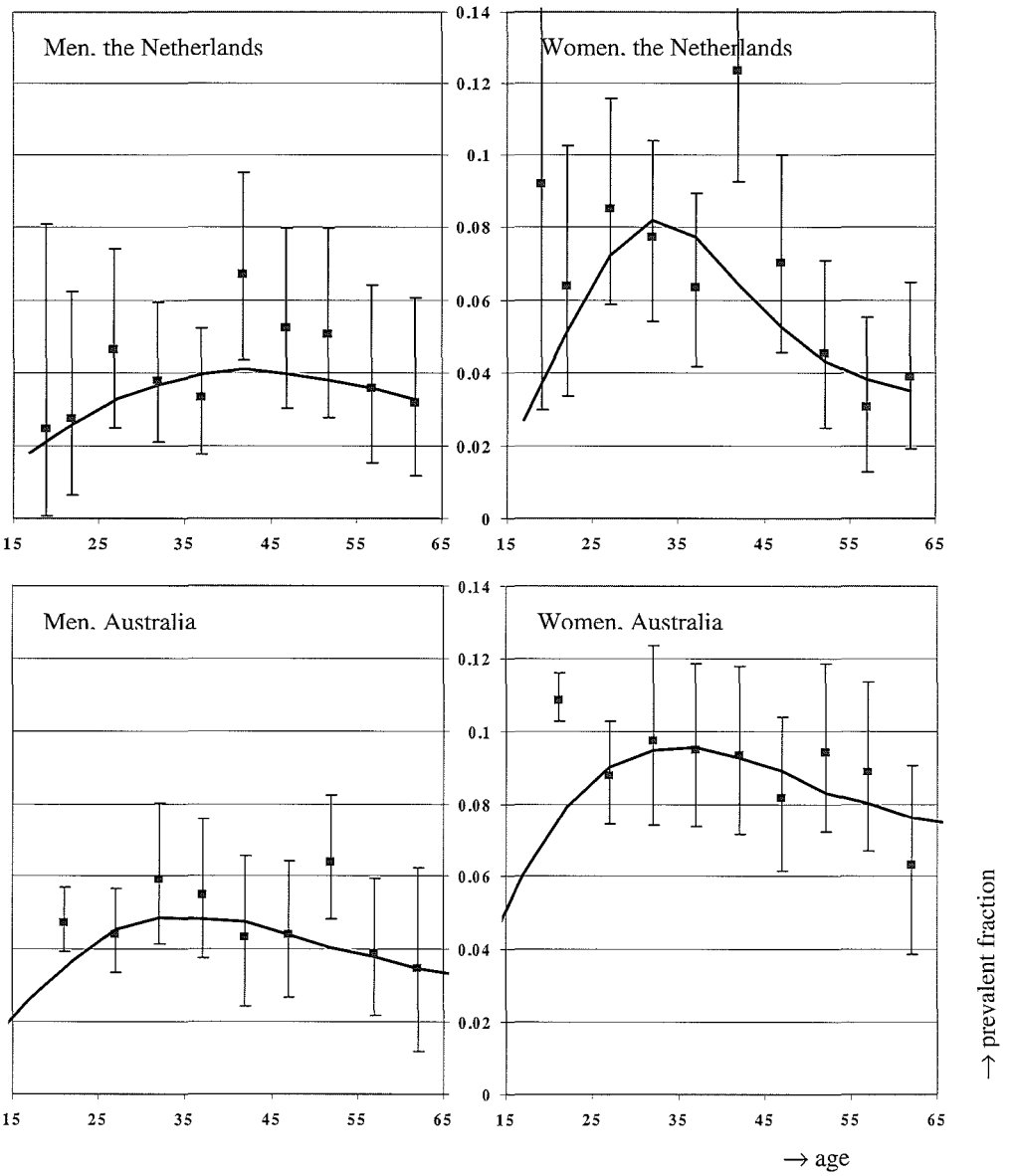
The model provides age-specific and a cohort estimate of lifetime prevalence. To enable comparison with the survey data we calculated a cross-sectional estimate averaging the age-specific model results for ages 15-64, weighting for the relative population size of the respective age groups in the Dutch 1996 (Statistics Netherlands) and Australian 1997 (Australian Bureau of Statistics) population.

We examined the effect of excess mortality of persons with a history of depression. Using a relative risks on total mortality (RR) we adjusted the survival curve for people with one or more episodes. We recalculated the fitted model using a RR of 1.81 that was reported in a meta-analysis of community studies [30], and an arbitrary higher value of 4.

Results

Figure 4.1 depicts the Dutch and Australian 12-month prevalence data by age and their 95% confidence intervals (95%CI); figure 4.2 the 1-month data. For all age groups the prevalence data are higher for women than for men, an observation that is consistently reported in the literature [31]. The rates fluctuate with age, with the 1-month prevalence having a similar, but lower, pattern than the 12-month prevalence. Between the two countries, the age-patterns are comparable, although the rates in Australia are on average higher than in the Netherlands. This does not necessarily relate to a true difference in the prevalence between these countries; differences in the version of the CIDI, DSM and the use of a lifetime or 12-months questionnaire as reference point may also cause such differences. Moreover, the 95% confidence intervals (95%CI) around the data are wide.

Figure 4.1. **Twelve-month prevalence of Major Depression for the Netherlands and Australia: model outcome and data for men and women by age.**



Figures 4.1 and 4.2 also show the modelled prevalence. For most age groups modelled prevalence falls within the 95%CI of the data. For the Netherlands, the data at age 40-45 are not reproduced well, while for Australia the data at age 50-55 and the 12-month data for the youngest age group are markedly higher than the model outcome.

In table 4.1 and 4.2 we give the model results for, respectively, mean duration and number of episodes and compare these to findings from community studies. Modelled mean episode duration is approximately 24 weeks, except for Australian men where it is 5-6 weeks longer. These estimates agree well with the number of weeks reported in several community surveys (Table 4.1), except for those reported by Kendler et al. and Spijker et al. [32, 33]. The mean number of episodes predicted by the model is around 7-8, but 13 for Australian women (Table 4.2). Our estimates are similar to the results from two community surveys that reported the number of depressive episodes (Table 4.2).

For both countries the model estimates that approximately 30% of men and 40% of women experience one or more episodes of MD during their entire lifetime (Table 4.3). In a cross-section of the modelled population aged 15-65, this is equivalent to a lifetime prevalence of 20% and 30%, respectively, which is higher than reported in recent community surveys (Table 4.3). The Dutch NEMESIS data are on average 38% lower than the cross-sectional model estimates ({model-data}/model*100%). Lifetime prevalence by age is shown for the Netherlands in Figure 4.3 (model estimates follow a similar pattern for Australia, survey data are not available). The model estimates increase with age until they reach a plateau, while the NEMESIS increase to a lower level and decline after age 45-50. The difference between the data and the model increases with age, with the data being approximately 70% lower at age 60-64. Figure 4.3 also depicts the results of adjusting for excess mortality from depression (a similar pattern was found for Australia). As can be seen from this graph, this had little effect on the model estimates.

Figure 4.2. **One-month prevalence of depression for the Netherlands and Australia: model outcome and data for men and women by age.**

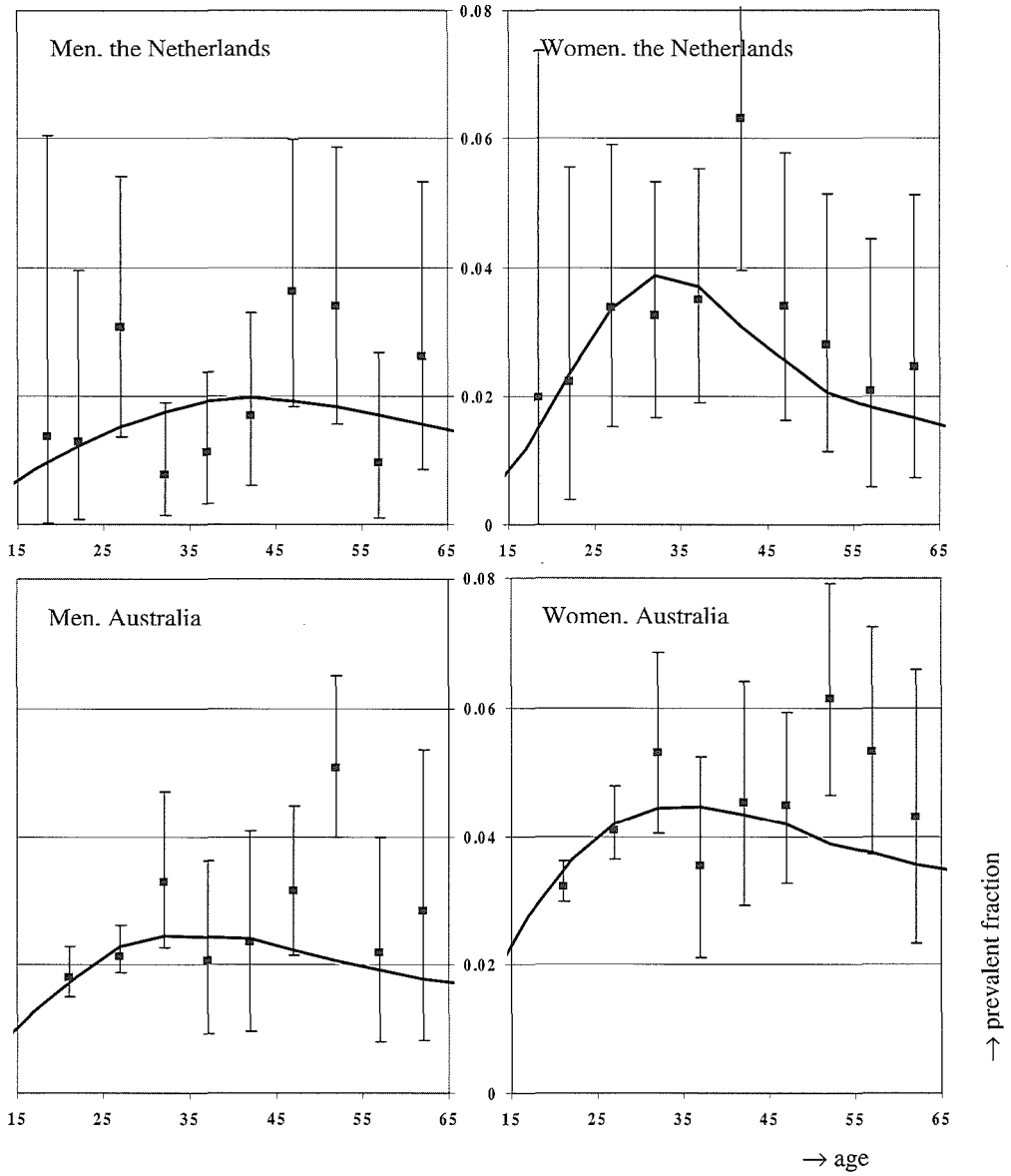


Table 4.1. **Mean episode duration (in weeks): model results and literature.**

Study	Men	Women	Both sexes
Modelling the Netherlands	24.6	23.8	24.1
Modelling Australia	29.5	24.4	25.8
ECA (USA) [28, 34]	26	27	*
NCS (USA) [11, 28]	*	*	23
Lewinsohn et al. (adolescents) [35]	*	*	26
Kendler et al. (twins) [32]	*	*	12
Spijker et al. (the Netherlands) [33]	*	*	36

ECA: Epidemiologic Catchment Area
NCS: National Comorbidity Survey
*: these data are not reported

Table 4.2. **Mean number of episodes: model results and literature.**

Study	Men	Women	Both sexes
Modelling the Netherlands	6.8	7.7	7.3
Modelling Australia	8.0	12.8	11
NCS [3]	*	*	10.8
Andrews et al. (twins) [36]	*	*	8

*: these data were not reported

Table 4.3. **Lifetime prevalence (percentage): model results and literature.**

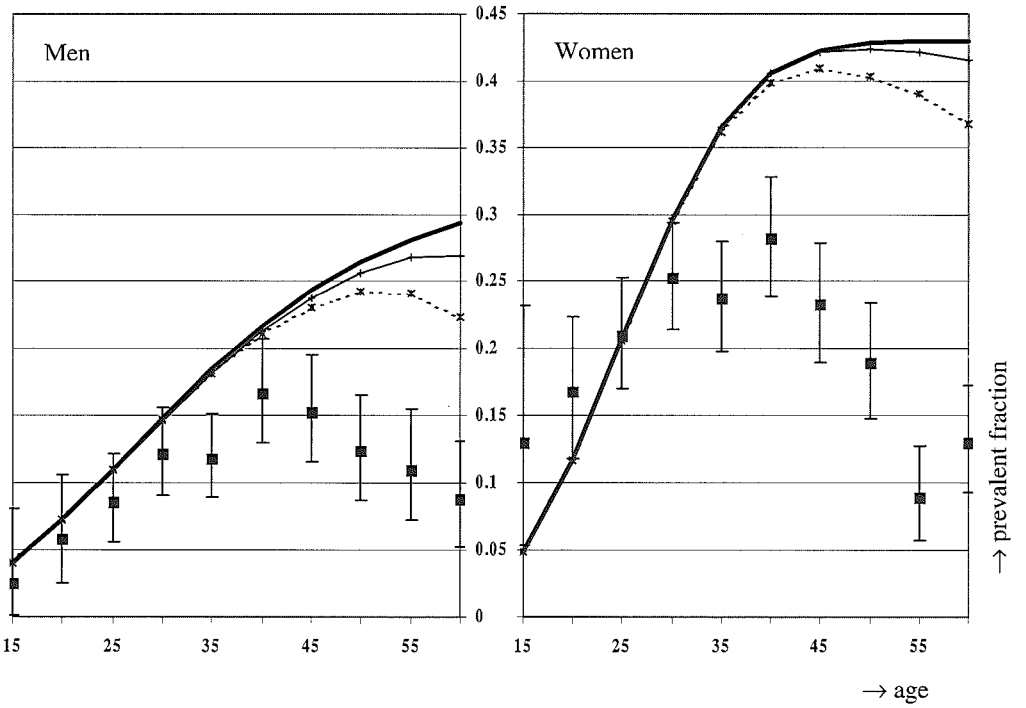
Study	Men	Women	Both sexes
Modelling the Netherlands (cohort)	29.8	42.4	36.2
Ditto (cross-section)	17.8	32.2	25.1
NEMESIS [10]	10.9	20.1	15.4
Modelling Australia (cohort)	27.2	43.4	35.4
Ditto (cross-section)	20.1	23.4	27.4
ECA [13]	*	*	5.8
NCS [11]	12.7	21.3	17.1

*: these data were not reported

Discussion

We estimated lifetime prevalence of major depression indirectly using a modelling approach and found that approximately 30% of men and 40% of women suffer from one or more episodes of MD during their life in the Netherlands and Australia. The cross-sectional equivalent of this (20% and 30%, respectively) is higher than reported in community surveys. The Dutch survey data (NEMESIS) are 38% lower than the model estimates, indicating a potentially large recall bias. The model replicated the data to which it was fitted well. The plausibility of the model is supported by the similarities between the outcomes of the two countries and a comparison to literature findings.

Figure 4.3. **Lifetime prevalence of depression for the Netherlands: data and model outcome by age.**



The model outcomes (prevalence, average duration, number of episodes) are mostly similar between the two countries and, apart from the well-described sex difference in the prevalence of MD, also between sexes. Two exceptions are the duration of episodes in Australian men and their number in Australian women. These higher estimates can be explained by the fact that the 1- and 12-month prevalence data were higher in the Australian than in the Dutch survey. To fit to these higher data the model can increase its current prevalence estimates in three ways: by increasing the number of susceptibles (i.e. lifetime prevalence), the duration of episodes, or their number. Episode duration was fixed, because we calculated it from 1- and 12-month prevalence. In Australian men, the ratio between these two variables was smaller than in the other three calculations, explaining the higher estimate. For Australian women, with an estimate of episode duration similar to the Dutch estimates, the only remaining variable to change is the number of episodes, as we reported the fit with the lowest lifetime prevalence.

The model-generated results are (mostly) in agreement with data from the international literature. The estimated mean duration of episodes of around 24 weeks is similar to results from several community studies in the USA [28, 35]. The higher estimate of the Dutch study may be due to the inclusion of sub-threshold cases and of dysthymic patients [33], the lower estimate of Kendler et al. [32] to the exclusion of persons with episodes lasting over a year. Also the number of episodes is comparable to the literature. This agreement with the literature contributes to the plausibility of the model, but more information, for example on the number of episodes encountered lifetime, will help to further pinpoint the model results.

The lifetime prevalence estimates are the only results that differ from the international literature. We have presented here the lowest estimates of lifetime prevalence that are consistent with the 1- and 12-month prevalence data from NEMESIS and the Australian survey. The lower lifetime prevalence figures reported by NEMESIS are thus inconsistent with the current prevalence data from the same study. There are two major sources for this inconsistency. Firstly, the NEMESIS estimates are not based on completed life-courses. It is not possible to obtain such information in a cross-sectional setting. Secondly, problems in the recall of symptoms and their timing will lower the number of reported lifetime cases in NEMESIS. After correcting the model estimates for a cross-sectional setting the NEMESIS data are still 38% lower than these

estimates, a difference that increases with age. Recall bias may explain this underestimation. Taking into account that the average time of recall in the survey will be smaller than the 25 years studied by Andrews et al. our 38% is not far off from the 50% underestimation reported by Andrews et al. [21].

The increased risk of mortality in persons with a history of MD can not explain a large part of this underestimation. A recent meta-analysis estimated the RR of mortality to be around 1.81 [30], but even taking into account a much higher RR of four hardly changed the average percentage of underestimation (35% at RR=4). The effect is only seen at the older ages where lifetime prevalence decreases. Thus, although excess mortality decreases the underestimation at higher ages, across all ages the underestimation of the data remains approximately similar.

One point that may be seen as a drawback of this study is that we did not include trends over time into the model. Several authors have suggested that incidence of (first episodes of) MD has increased over time [37]. A trend of increasing incidence would cause the lifetime prevalence rates at a certain age to be higher in younger cohorts than in older ones and may be an explanation for the observed decrease in the NEMESIS lifetime rates after age 45, which is not seen in our model. However, such a trend should have only little influence on our model estimates, as it also effects the level of current prevalence by age. Thus, by fitting the model to current prevalence data, the effect of an increase in incidence is already partially included in the model. We therefore expect the trend in incidence, like excess mortality, not to detract from our higher overall estimates of lifetime MD prevalence.

Conclusion

Based on Dutch and Australian 1-month and 12-month prevalence data, we estimated that approximately 30% of men and 40% of women suffer from one or more episodes of MD during their life. Lifetime rates observed in cross-sectional surveys are much lower. Two important reasons for this are that these surveys ignore cases that become incident after the time of the survey and that self-reported symptoms and their timing may be biased by recall problems. We showed the Dutch data to be underestimated through recall bias by at least 35%. We expect excess mortality and trends in incidence to have little influence on this estimate. The large recall problem poses questions to the measurement of lifetime prevalence in cross-sectional surveys.

Acknowledgement

This study was supported by the Netherlands Institute of Health Sciences, the Netherlands Organization for Scientific Research and the Erasmus University Trustfund.

References

1. Broadhead WE, Blazer DG, George LK, Tse CK. Depression, disability days, and days lost from work in a prospective epidemiologic survey. *Jama* 1990;264(19):2524-8.
2. Bijl RV, Ravelli A. Current and residual functional disability associated with psychopathology: findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Psychol Med* 2000;30(3):657-68.
3. Kessler RC, Zhao S, Blazer DG, Swartz M. Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *J Affect Disord* 1997;45(1-2):19-30.
4. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *Jama* 1989;262(7):914-9.
5. Ustun TB, Sartorius N. *Mental Illness in General Health Care: an International Study*. London: John Wiley and Sons; 1995.
6. Sprangers MA, de Regt EB, Andries F, van Agt HM, Bijl RV, de Boer JB, et al. Which chronic conditions are associated with better or poorer quality of life? *J Clin Epidemiol* 2000;53(9):895-907.
7. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge: Harvard University Press; 1996.
8. Mathers CD, Stein C, Ma Fat D, Chalapati R, Inoue M, Tomijima N, et al. *Global Burden of Disease 2000: version 2 methods and results*. World Health Organization, 2002: http://www3.who.int/whosis/discussion_papers/discussion_papers.cfm?path=whosis,burden_statistics,discussion_papers&language=english. Accessed 2003.
9. Andrews G, Henderson S, Hall W. Prevalence, comorbidity, disability and service utilisation. Overview of the Australian National Mental Health Survey. *Br J Psychiatry* 2001;178:145-53.
10. Bijl RV, Ravelli A, van Zessen G. Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc Psychiatry Psychiatr Epidemiol* 1998;33(12):587-95.

11. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry* 1994;51(1):8-19.
12. Offord DR, Boyle MH, Campbell D, Goering P, Lin E, Wong M, et al. One-year prevalence of psychiatric disorder in Ontarians 15 to 64 years of age. *Can J Psychiatry* 1996;41(9):559-63.
13. Regier DA, Boyd JH, Burke JD, Jr., Rae DS, Myers JK, Kramer M, et al. One-month prevalence of mental disorders in the United States. Based on five Epidemiologic Catchment Area sites. *Arch Gen Psychiatry* 1988;45(11):977-86.
14. Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC. Long-term reliability of diagnosing lifetime major depression in a community sample. *Arch Gen Psychiatry* 1986;43(5):435-40.
15. Fendrich M, Weissman MM, Warner V, Mufson L. Two-year recall of lifetime diagnoses in offspring at high and low risk for major depression. The stability of offspring reports. *Arch Gen Psychiatry* 1990;47(12):1121-7.
16. Prusoff BA, Merikangas KR, Weissman MM. Lifetime prevalence and age of onset of psychiatric disorders: recall 4 years later. *J Psychiatr Res* 1988;22(2):107-17.
17. Wade TD, Bulik CM, Kendler KS. Reliability of lifetime history of bulimia nervosa. Comparison with major depression. *Br J Psychiatry* 2000;177:72-6.
18. Warshaw MG, Klerman GL, Lavori PW. Are secular trends in major depression an artifact of recall? *J Psychiatr Res* 1991;25(3):141-51.
19. Wittchen HU, Burke JD, Semler G, Pfister H, Von Cranach M, Zaudig M. Recall and dating of psychiatric symptoms. Test-retest reliability of time-related symptom questions in a standardized psychiatric interview. *Arch Gen Psychiatry* 1989;46(5):437-43.
20. Angst J. Epidemiology of depression. *Psychopharmacology (Berl)* 1992;106 Suppl:S71-4.
21. Andrews G, Anstey K, Brodaty H, Issakidis C, Luscombe G. Recall of depressive episode 25 years previously. *Psychol Med* 1999;29(4):787-91.
22. Bijl RV, van Zessen G, Ravelli A, de Rijk C, Langendoen Y. The Netherlands Mental Health Survey and Incidence Study (NEMESIS): objectives and design. *Soc Psychiatry Psychiatr Epidemiol* 1998;33(12):581-6.
23. World Health Organization. *Composite International Diagnostic Interview (CIDI), version 1.0*. Geneva: World Health Organization; 1990.
24. Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic

Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988;45(12):1069-77.

25. APA. *Diagnostic and statistical manual of mental disorders : DSM-III-R.* 3rd , rev. ed. Washington, DC: American Psychiatric Association; 1987.
26. Radhakrrishna Rao C. *Linear statistical inference and its applications.* 2nd ed. New York: Wiley; 1973.
27. Habbema JD, De Vlas SJ, Plaisier AP, Van Oortmarssen GJ. The microsimulation approach to epidemiologic modeling of helminthic infections, with special reference to schistosomiasis. *Am J Trop Med Hyg* 1996;55(5 Suppl):165-9.
28. Ustun TB, Kessler RC. Global burden of depressive disorders: the issue of duration. *Br J Psychiatry* 2002;181:181-3.
29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical recipes in C: the art of scientific computation.* 2 ed. Cambridge: Cambridge University Press; 1995.
30. Cuijpers P, Smit F. Excess mortality in depression: a meta-analysis of community studies. *J Affect Disord* 2002;72(3):227-36.
31. Kessler RC. Epidemiology of women and depression. *J Affect Disord* 2003;74(1):5-13.
32. Kendler KS, Walters EE, Kessler RC. The prediction of length of major depressive episodes: results from an epidemiological sample of female twins. *Psychol Med* 1997;27(1):107-17.
33. Spijker J, de Graaf R, Bijl RV, Beekman AT, Ormel J, Nolen WA. Duration of major depressive episodes in the general population: results from The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Br J Psychiatry* 2002;181:208-13.
34. Eaton WW, Anthony JC, Gallo J, Cai G, Tien A, Romanoski A, et al. Natural history of Diagnostic Interview Schedule/DSM-IV major depression. The Baltimore Epidemiologic Catchment Area follow-up. *Arch Gen Psychiatry* 1997;54(11):993-9.
35. Lewinsohn PM, Clarke GN, Seeley JR, Rohde P. Major depression in community adolescents: age at onset, episode duration, and time to recurrence. *J Am Acad Child Adolesc Psychiatry* 1994;33(6):809-18.
36. Andrews G. Should depression be managed as a chronic disease? *Bmj* 2001;322(7283):419-21.
37. Simon GE, VonKorff M, Ustun TB, Gater R, Gureje O, Sartorius N. Is the lifetime risk of depression actually increasing? *J Clin Epidemiol* 1995;48(9):1109-18.

PART B

Tailoring health status valuations to the epidemiology

5

Levels of disability in Major Depression
Findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS)

M.E. Kruijshaar, N. Hoeymans, R.V. Bijl, J. Spijker, M.L. Essink-Bot. Levels of disability in Major Depression. Findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Journal of Affective Disorders* 2003; 77 (1): 53-64.

Summary

Background

Information on the distribution of disability associated with Major Depression (MD) across different groups of patients is of interest to health policy and planning. We examined the associations of severity and type (a single or recurrent episode) of MD with disability in a Dutch general population sample.

Methods

We used data from the first wave (1996) of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). MD 'severity' and 'type' were diagnosed with the help of the Composite International Diagnostic Interview according to DSM-III-R criteria. SF-36 scores, days ill in bed and days absent from work were taken as indicators of disability. The differences in these variables were studied by means of variance and regression analysis.

Results

Recurrent MD was found not to be associated with more disability than single episode MD. Higher 'severity' classes were associated with more disability. However, the degree of disability between 'moderate' and 'severe' MD differed only very slightly. The difference in disability between non-depressed and mildly depressed individuals had a larger effect than between each successive pair of 'severity' classes.

Conclusions

Three groups of MD can be distinguished based on the associated degree of disability: 'mild', 'moderate to severe' and 'severe with psychotic features'. In the future, these groups can be used to describe the distribution of disability in the depressed population. The marked difference between 'mild' MD and no MD suggests that 'mild' cases should be considered relevant.

Background

Recent community surveys reveal an alarmingly high prevalence of Major Depression (MD) in several developed countries [1-4]. In addition, the impact of MD on the daily functioning of the individual is strong [5-7], and the limitations in well-being and functioning of patients have been shown to be equal to or greater than those of several major chronic medical conditions [8-11]. The distribution of disability across the depressed population, however, has received little attention. Identifying differences in the extent of disability between sub-groups of depressed patients may prove useful to health policy and planning and may help to interpret the high prevalence figures.

Cases of depression are often classified into sub-groups according to the criteria of the American Psychiatric Association (APA) to facilitate treatment planning in the clinical situation. The third Revised Version of the Diagnostic and Statistical Manual of mental disorders of the APA (DSM-III-R) [12] classifies episodes of MD into two 'types', namely 'single' and 'recurrent'. Furthermore, an episode is labelled by level of 'severity': 'mild', 'moderate', 'severe', 'severe with psychotic features' and MD of 'unspecified severity'. 'Severity' is defined according to the DSM-III-R by the number of depressive symptoms and the limitations in social, emotional and professional functioning that these symptoms cause [12]. The more severe the label assigned to an episode, the greater the likelihood that the episode will be associated with more disability, although empirical evidence to support this is lacking. Furthermore, in clinical practice it is often assumed that recurrent MD has a larger impact on functioning than single MD, but this has not been documented in the literature.

It has been suggested that, despite the use of standardised diagnostic criteria, milder cases of MD are identified in community surveys compared to clinical settings [13]. If less disability is associated with mild cases diagnosed in a community setting, compared to those in the clinics, the question is whether they are still of public health concern. This could be investigated by comparing the extent of disability associated with 'mild' MD in community surveys to non-MD.

Our study focussed on the following assumptions: 1) that recurrent MD is associated with more disability than single MD, 2) that higher 'severity'

classes are associated with more disability, and 3) that individuals with 'mild' MD experience distinctly more disability than non-depressed individuals.

To study these assumptions, we analysed how disability differed according to MD 'type' and 'severity', and between mildly depressed and non-depressed individuals. We studied this for the Dutch population using data from the Netherlands Mental Health Survey and Incidence Study (NEMESIS) [14], in which MD 'type' and 'severity' were diagnosed using the Composite International Diagnostic Interview (CIDI) [15]. Disability was defined as limitations in the physical, psychological and social domains of functioning.

Materials and methods

Study sample

Data were collected in the Netherlands Mental Health Survey and Incidence Study (NEMESIS). The methods used have been described elsewhere by Bijl et al. [4, 14]. Briefly, the survey was based on a three-stage, stratified random sample drawn from the Dutch general adult population, aged 18 to 64. An initial sample was drawn from a population of 90 Dutch municipalities, stratified by urbanicity and province, the next from a population of private households (addresses from post office registers) and a third from a population of individuals (most recent birthday). The survey was prospective: the interviews were taken in 1996, 1997 and 1999. We used cross-sectional unweighted data obtained on individuals diagnosed with MD in 1996.

The interviews were spread over the entire year 1996 to compensate for seasonal influences. A maximum of ten attempts were made at various times and on various days to contact interviewees who were initially unavailable. This increased the response rate, and reduced selection bias due to short-term institutionalisation. The response rate for the sample was 69.7%, yielding a study population of 7076 persons, of whom 439 had been diagnosed with MD within the past year and 204 within the past month. Investigation of the non-responders (i.e. refusals) revealed no evidence of selectivity (43% of non-responders could be assessed). The study sample (n=7076) was shown to be representative of the Dutch adult population in terms of gender, civil status, and degree of urbanisation of the place of residence. The 18-24 year age group was slightly underrepresented.

Instruments/ variables measured

The Composite International Diagnostic Interview (CIDI; version 1.1, computerised) [4, 14-16] was the diagnostic interview used to establish the diagnoses of MD and to classify MD 'type' and 'severity' based on the DSM-III-R [12]. The CIDI is a structured interview developed by WHO. It can be administered by trained non-clinicians. The validity and reliability of this instrument have been shown to be good for all psychiatric diagnoses including MD [17]. However, its capacity to classify MD by 'type' and 'severity' has not been validated.

Disability has been operationalised in different ways [18-21]. In the International Classification of Impairments, Disabilities and Handicaps (ICIDH, [18]) disability is restricted to disturbances at the level of the individual, as distinct to the level of the human organism (impairment) and in the societal context (handicap). In the International Classification of Functioning, Disability and Health (ICF, [19]) that has recently succeeded it, disability is, however, used in a broader sense, covering limitations in all levels of functioning. We used disability in its broader meaning. We defined disability as limitations in physical, psychological and social functioning. Disability was measured using the Short-Form-36 Health Survey (SF-36) [22, 23], an instrument reported to provide good reliability and validity [23]. Furthermore, it has proven to be able to distinguish different severity levels of MD in the elderly [24]. The 36 items of the SF-36 are grouped into eight health scales: physical functioning, role limitations due to physical problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems and mental health. The scoring algorithm recommended by Ware & Sherbourne was adhered to [22], in which scores are distributed between 0 and 100, the latter representing optimal health.

We included two other indicators of disability to reflect some of the economic consequences of depression as well: the number of days spent in bed due to psychiatric, drug or alcohol related problems (days ill in bed), and the time missed from work due to these problems (absence days). In both cases, the data collected referred to the past 12 months (measured in number of days) and to the past month (five classes: 0 days, 1-3 days, 4-7 days, >1week, <1 month, the whole time).

We included sex, age (in 10-year age groups) and socio- economic status (SES) as demographic confounders in our analyses. We used educational attainment (four levels) as a proxy for SES. Other confounders included were adverse early life experiences, the presence of chronic somatic conditions, and the presence of other psychiatric disorders. These variables were found to be associated with MD in an earlier study [7] and may affect an individual's functioning. The variable 'adverse early life experiences' consisted of self-reported emotional neglect, psychological abuse, physical abuse or unwanted sexual approaches on more than one occasion prior to age 16. The presence of chronic somatic conditions during the previous 12 months was assessed by a list of 31 chronic somatic conditions, based on the Health Interview Survey of Statistics Netherlands [25]. Only those somatic conditions that were reported to be treated or monitored by a doctor were included in the analyses. 'Other psychiatric disorders' included anxiety disorders, substance dependence or abuse (including alcohol), dysthymia, bipolar disorder, eating disorders and schizophrenia, as diagnosed by the CIDI.

The CIDI provides diagnoses of MD (and presence of other psychiatric disorders) in the past month and year (and lifetime). Because we were not only interested in the extent of the problems during an episode, but also after this episode (residual disability), we used the sample of individuals diagnosed with MD in the past year in our analyses. The variables days ill in bed and absence days, and the two comorbidity confounders that refer to the past year were used. The SF-36 scores, however, reportedly measure the health situation during the past month. Although, the sample of individuals prevalent in the past month corresponds better with the SF-36 data, there are two additional reasons why we did not use it. First, this does not improve the correspondence with the timings of 'severity' and disability. 'Severity' is assessed by the CIDI for a severe -or the most severe- period of an episode, which may easily have occurred before the last month as the average duration of an episode is 4-6 months. Second, the number of individuals depressed in the last month was too small to reveal significant results in the comparison of the separate 'severity' classes. To examine whether the relationships differ using this sample, we repeated the regression analyses (see below), using variables and confounders that refer to the last month (except presence of somatic disorders), and inspected the beta-coefficients.

Statistical analyses

We used chi-square analysis to study the association between 'type' and 'severity' of MD.

Analyses of variance (ANOVA) were performed to study the associations of 'type' and 'severity' with each of the ten indicators of disability. We constructed ten models in which 'type' and level of 'severity' constituted the independent variables and one of the disability indicators the dependent variable. The confounders were adjusted for in the analyses.

To test for a trend of increasing disability with recurrent 'type' and increasing 'severity', we used linear regression analyses, adjusting for the same confounders.

When 'severity' was found to be statistically significant in the ANOVA, we tested which successive pairs of severity classes differed from each other by means of ANOVA. We corrected for the same confounders, but 'type' was not included in the analysis. To assess whether statistically significant differences in these tests were meaningful, we estimated effect sizes (*d*) using Cohen's guidelines to interpret the size of the effect: $0.2<=d<0.5$ for a small, $0.5<=d<0.8$ for a moderate and $d>=0.8$ for a large effect [26].

In the same way, we compared mean SF-36 scores, days ill in bed, and absence days between individuals with 'mild' MD and without MD, and between those with MD and without MD. We also estimated effect sizes as described above. The statistical package used was SAS (version 6.12 [27]).

Results

The study population

The characteristics of the total sample of the NEMESIS survey, and of the individuals diagnosed with MD in the past year and month are shown in Table 5.1. The unweighted prevalence of MD in the last year in the total NEMESIS sample was 6.2%. The generally reported prevalence of MD weighted for sex, age, urbanicity and marital status is 5.8% (NEMESIS [4]). Compared to non-depressed individuals, depressed individuals were significantly more likely to be female, to report more adverse early life events, somatic disorders and to have a low level of educational attainment. Of individuals who were diagnosed with MD 69% was diagnosed with a non-affective psychiatric disorder, compared to

only 18% of non-depressed individuals, a significant difference. Depressed individuals reported significantly higher levels of disability: the mean SF-36 scores were lower (higher SF-36 scores indicate a better health status), whereas the mean number of days ill in bed and mean number of absence days were both higher. For most variables, these patterns seemed more pronounced for MD in the past month (Table 5.1).

Disability according to 'type' and 'severity'

Table 5.2 shows the mean SF-36 scores, mean number of days ill in bed and mean number of absence days (all unadjusted), broken down according to 'type'

Table 5.1. **Characteristics of the study population.**

Population:	Total population		MD in the last year		MD in the last month	
Number (prevalence in %)	7076	(100)	439	(6.2)	204	(2.9)
Confounders						
% Females	53.4		67.0		66.7	
Mean age in years (SD)	41.2	(12.2)	39.7	(11.0)	41.5	(11.1)
% Educational level: high	27.3		23.0		21.6	
% Adverse early life experience	28.7		56.3		59.3	
% Chronic somatic condition(s) (treated)	41.0		51.9		57.4	
% Psychiatric disorder(s) other than MD	21.1		69.0		74.5	
Disability indicators						
Mean SF-36 scores (sd)						
Physical functioning	91.3	(16.5)	85.6	(20.9)	81.2	(23.5)
Role limitations, physical	85.1	(30.5)	70.5	(40.0)	63.4	(42.8)
Bodily pain	84.8	(22.0)	73.6	(27.5)	68.0	(29.0)
General health perception	73.9	(18.2)	62.4	(22.0)	55.7	(22.4)
Vitality	71.1	(18.6)	51.1	(22.4)	40.3	(20.4)
Social functioning	89.3	(18.4)	71.4	(25.8)	62.5	(26.6)
Role limitations, emotional	91.9	(23.8)	61.3	(42.1)	42.6	(41.7)
Mental health	81.5	(15.3)	59.8	(22.3)	46.6	(19.8)
Mean # of days ill in bed (sd)	0.6	(8.3)	5.0	(23.2)	–[a]	
% with >0 days ill in bed	3.8		16.5		9.3	
Mean # of absence days (sd)	4.2	(31.1)	32.9	(81.9)	–[a]	
% with >0 absence days	6.6		33.7		17.6	

[a] The number of days in the last month were measured categorically (five levels), and no mean value could be calculated. SD, standard deviation

and 'severity', and according to 'type' alone. Inspecting this Table, it seems that mean SF-36 scores were somewhat lower (i.e. worse) for depressed individuals diagnosed with a single episode than for individuals with recurrent episodes of MD, and the mean number of absence days somewhat higher. Split up by 'severity' class, the scores for single 'severe' MD looked particularly worse for all ten indicators compared to recurrent 'severe' MD.

For higher 'severity' classes, it seems that the mean SF-36 scores were somewhat lower and the mean number of days ill in bed and absence days higher. This pattern, however, does not appear for 'moderate' and 'severe' MD. The scores of the latter two 'severity' classes seem to lie close together, as also appears from Figure 5.1, which shows the (unadjusted) mean scores of the indicators by 'severity'.

The mean scores of single and recurrent MD of 'unspecified severity' are also given in Table 5.2, but we excluded these groups of persons from further analysis because they could not be categorised accurately into one of the four other 'severity' classes.

Association of 'type' and 'severity'

Chi-square analysis of MD 'severity' versus 'type' revealed no significant relation ($p=0.486$); the distribution of cases across the 'severity' classes did not differ significantly between single and recurrent episode MD.

Association of 'type' and 'severity' with disability

In the analysis of variance we adjusted for sex, age, educational attainment, adverse early life experience, presence of chronic somatic conditions and presence of other psychiatric disorders. 'Type' of MD was not significantly different for any of the disability indicators. The 'severity' classes, on the other hand, differed significantly on all SF-36 scales, days ill in bed and absence days ($p<0.05$).

Direction of the association

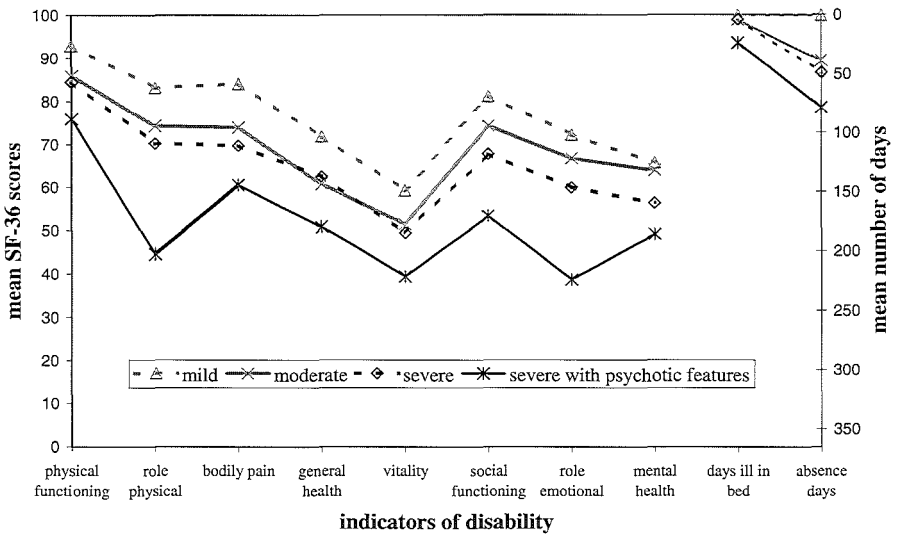
In the regression analysis we adjusted for the same confounders. There was no significant trend with 'type' of MD. A positive linear trend with 'severity' was found for all SF-36 scales, for the number of days ill in bed and for absence days ($p<0.05$).

Table 5.2. **SF-36 scores, number of days ill in bed and absence days according to 'type' and 'severity' class: means and sd.**

12-Month diagnosis	Disability indicator										Numbers observed
	SF-36[a]								Other		
	Physical functioning	Role limitations (physical)	Pain	General health Perception	Vitality	Social functio-ning	Role limitations (emotional)	Mental health	Days ill in bed	Absence days	
Single MD, total	83.9	67.9	73.1	60.8	50.8	70.9	60.8	58.7	4.8	37.0	240
	(22.8)	(44.0)	(28.0)	(22.9)	(23.4)	(26.9)	(42.6)	(24.1)	(23.3)	(89.4)	
Single 'mild'	92.2	85.5	82.1	70.8	62.1	85.3	78.0	67.6	0.2	0.3	50
	(14.0)	(29.1)	(22.3)	(18.1)	(19.2)	(18.2)	(33.4)	(20.1)	(1.0)	(1.4)	
Single 'moderate'	85.4	70.4	76.3	57.8	51.2	76.2	68.0	66.5	6.0	48.8	49
	(21.2)	(41.0)	(27.7)	(21.8)	(22.7)	(24.4)	(40.2)	(19.9)	(26.9)	(87.1)	
Single 'severe'	81.4	67.4	68.3	59.3	46.5	64.4	57.5	52.0	3.8	53.9	70
	(26.1)	(40.3)	(29.5)	(23.1)	(23.7)	(26.7)	(43.1)	(25.3)	(22.0)	(109.3)	
Single psychotic	75.3	44.2	62.8	51.3	39.2	52.7	34.4	48.5	18.6	81.3	30
	(24.6)	(44.9)	(30.9)	(26.6)	(23.0)	(27.7)	(42.4)	(25.0)	(43.5)	(129.5)	
Single not specified	82.4	61.6	74.1	61.3	52.1	71.7	56.1	57.4	0.6	8.3	41
	(23.9)	(43.0)	(27.0)	(23.1)	(23.2)	(28.3)	(45.0)	(24.8)	(2.0)	(39.6)	
Recurrent MD, total	87.7	73.6	74.0	64.3	51.4	72.0	62.0	61.1	5.3	28.1	199
	(18.1)	(38.8)	(27.1)	(20.8)	(21.3)	(24.4)	(41.6)	(19.9)	(23.3)	(71.8)	
Recurrent 'mild'	93.6	80.9	86.3	72.8	56.2	76.9	66.0	63.6	0.7	1.0	47
	(8.5)	(35.4)	(17.9)	(18.0)	(21.4)	(20.4)	(40.2)	(18.8)	(2.6)	(2.8)	
Recurrent 'moderate'	86.4	77.9	71.6	63.6	51.6	72.2	65.4	61.5	3.2	30.1	52
	(20.3)	(35.6)	(29.3)	(17.1)	(21.3)	(24.2)	(42.8)	(20.3)	(17.0)	(72.6)	
Recurrent 'severe'	88.5	74.1	71.5	66.5	53.1	71.9	63.0	61.9	4.2	42.5	54
	(16.5)	(38.8)	(25.3)	(19.1)	(20.1)	(26.0)	(39.7)	(20.3)	(18.2)	(86.4)	
Recurrent psychotic	76.8	45.0	57.0	50.0	39.8	54.3	45.0	49.8	31.2	74.7	20
	(22.1)	(44.1)	(31.7)	(27.1)	(21.1)	(26.0)	(44.9)	(21.1)	(55.9)	(111.6)	
Recurrent not speci-fied	86.0	73.1	75.2	56.7	47.7	76.5	59.0	62.9	0.5	7.2	26
	(22.0)	(39.3)	(28.3)	(23.4)	(21.1)	(21.8)	(42.5)	(17.4)	(1.5)	(27.7)	
Total MD	85.6	70.5	73.6	62.4	51.1	71.4	61.3	59.8	5.0	32.9	439
	(20.9)	(40.0)	(27.5)	(22.0)	(22.4)	(25.8)	(42.1)	(22.3)	(23.2)	(81.9)	

Standard deviations (sd) are shown in brackets. Psychotic, 'severe MD with psychotic features'; not specified, MD of 'unspecified severity'.
[a] The SF-36 scales are scored from 0-100, 100 representing optimal health.

Figure 5.1. **Unadjusted mean disability scores by 'severity' class.**



Comparison of 'severity' classes

We compared disability between the adjoining pairs of 'severity' classes, adjusting for the same confounders, and studied the relevance of significant differences, using effect size estimation. In Table 5.3, the means and effect sizes are reported; effect sizes of non-significant differences are not interpreted. 'Mild' and 'moderate' MD differed significantly in four SF-36 scales and in absence days. The effect sizes were small (4x) and moderate (1x), respectively. 'Moderate' and 'severe' MD differed only in mental health, a difference with a small effect size. Between 'severe' MD and 'severe with psychotic features' four SF-36 scales differed significantly, and also days ill in bed. The effect sizes were small (4x) and moderate (1x), respectively.

Comparison with non-depressed persons

Table 5.4 shows the results of the comparisons between non-depressed and depressed individuals and between non-depressed and 'mildly' depressed subjects, adjusted for confounders. As a group, depressed individuals differed

from non-depressed individuals on all disability indicators, although the effect of the difference in physical functioning was not meaningful. 'Mild' MD on the other hand differed significantly from non-MD in vitality, social functioning, role limitations due to emotional problems and mental health; four SF-36 scales that reflect psychological and social functioning. The effect sizes were moderate (1x), small (1x) and large (2x), respectively.

The effects of the differences between 'mild' MD and non-MD were larger than the differences between the subsequent 'severity' classes reported above, which had a small effect nine out of 11 times and never had a large effect. In addition to the differences in indicators of psychological and social functioning between 'mild' MD and non-MD, differences also emerged with respect to indicators reflecting physical functioning and economical consequences (physical functioning, role limitations due to physical problems, pain, general health and days ill in bed and absence days) between the successive 'severity' classes.

Regression analyses based on last month prevalence

We repeated the regression analyses using the sample of individuals diagnosed with MD in the past month (204 individuals). 'Type' remained insignificant. According to 'severity' disability differed less significantly in nine of the indicators to the extent that no significant differences could be observed in two. The regression coefficients were nevertheless larger, except for role functioning due to emotional and due to physical problems.

Discussion

We compared disability as measured by the SF-36, numbers of days ill in bed and number of absence days, between MD 'type' (single and recurrent) and 'severity', diagnosed according to the DSM-III-R. Recurrent episode MD was not found to be associated with more disability than single episode MD. Higher 'severity' classes on the other hand ('mild', 'moderate', 'severe' and 'severe MD with psychoses') were associated with increasing levels of disability, although not all classes differed significantly from each other. 'Mild' MD showed distinctly worse scores than non-MD in indicators that represent psychological and social functioning.

Table 5.3. **Comparison of disability indicators between successive 'severity' classes: raw mean scores, effect sizes and their interpretation.**

Variable	Means				Effect size	Interpre-tation
	Mild	Moderate	Severe	Psychotic		
Mild versus moderate MD						
SF-36						
Physical functioning	92.89	85.94			0.32	small
Role limitations (physical)	83.25	74.26			0.15	ns
Pain	84.12	73.89			0.28	small
General health	71.75	60.74			0.47	small
Vitality	59.23	51.44			0.29	small
Social functioning	81.20	74.16			0.21	ns
Role limitations (emotional)	72.16	66.67			0.17	ns
Mental health	65.65	63.92			0.02	ns
Other						
Days ill in bed	0.41	4.56			0.21	ns
Absence days	0.60	39.08			0.76	moderate
Moderate versus severe MD						
SF-36						
Physical functioning		85.94	84.51		0.07	ns
Role limitations (physical)		74.26	70.33		0.15	ns
Pain		73.89	69.70		0.12	ns
General health		60.74	62.48		0.14	ns
Vitality		51.44	49.39		0.11	ns
Social functioning		74.16	67.66		0.24	ns
Role limitations (emotional)		66.67	59.89		0.06	ns
Mental health		63.92	56.33		0.29	small
Other						
Days ill in bed		4.56	3.98		0.07	ns
Absence days		39.08	48.88		0.05	ns
Severe versus severe MD with psychotic features						
SF-36						
Physical functioning			84.51	75.90	0.38	small
Role limitations (physical)			70.33	44.50	0.42	small
Pain			69.70	60.50	0.09	ns
General health			62.48	50.82	0.42	small
Vitality			49.39	39.40	0.33	ns
Social functioning			67.66	53.30	0.36	ns
Role limitations (emotional)			59.89	38.67	0.44	small
Mental health			56.33	49.04	0.17	ns
Other						
Days ill in bed			3.98	23.71	0.67	moderate
Absence days			48.88	78.59	0.17	ns

Effect sizes are interpreted according to Cohen's guidelines: with cut-off values of 0.2, 0.5 and 0.8 for small, moderate and large effects, respectively. Effect sizes are adjusted for sex, age, educational attainment, adverse early life experience, presence of chronic somatic conditions and presence of other psychiatric disorders.
ns, the difference was not significant and the effect size, therefore, not interpreted.

Table 5.4. **Comparison of disability indicators between 'mild' MD and non-MD and between MD and non-MD: raw mean scores, effect sizes and their interpretation.**

Variable	Means			Effect size	Interpre-tation
	No MD	MD	Mild MD		
No MD versus MD					
SF-36					
Physical functioning	91.72	85.61		0.18	no effect
Role limitations (physical)	86.11	70.49		0.30	small
Pain	85.50	73.55		0.28	small
General health	74.64	62.38		0.39	small
Vitality	72.40	51.05		0.84	large
Social functioning	90.45	71.40		0.75	moderate
Role limitations (emotional)	93.91	61.34		1.16	large
Mental health	82.93	59.79		1.13	large
Other					
Days ill in bed	0.33	5.01		0.45	small
Absence days	2.34	32.92		0.84	large
No MD versus mild MD					
SF-36	91.72		92.89	0.12	ns
Physical functioning	86.11		83.25	0.02	ns
Role limitations (physical)	85.50		84.12	0.03	ns
Pain	74.64		71.75	0.05	ns
General health	72.40		59.23	0.55	moderate
Vitality	90.45		81.20	0.37	small
Social functioning	93.91		72.16	0.85	large
Role limitations (emotional)	82.93		65.65	0.95	large
Mental health					
Other					
Days ill in bed	0.33		0.41	0.07	ns
Absence days	2.34		0.60	0.20	ns

Effect sizes are interpreted according to Cohen's guidelines: with cut-off values of 0.2, 0.5 and 0.8 for small, moderate and large effects, respectively. Effect sizes are adjusted for sex, age, educational attainment, adverse early life experience, presence of chronic somatic conditions and presence of other psychiatric disorders.
ns, the difference was not significant and the effect size, therefore, not interpreted.

In contrast to what we expected, our findings do not support the general idea that episodes of recurrent MD are associated with more disability than a single episode. This finding is based on the insignificant association of 'type' in the ANOVA and the insignificant chi-square analysis. The latter indicates that individuals suffering from recurrent MD were not more likely to have a higher

'severity' label. One may argue against our finding that we did not test for the interaction between 'type' and 'severity'. To examine its effects we included it in additional analyses. It was not significant except for the mental health scale, indicating that disability varies similarly between the 'severity' classes for single and recurrent MD, except for the variation in mental health. Another demur might be that our dependent variables do not cover all aspects of disability, thereby concealing possible existing differences. Because at least all three domains of disability were represented in our variables, this, in our view, is not a major shortcoming. Also, the temporality of a diagnosis of a single episode of MD may be seen as a problem. This diagnosis may later change to recurrent MD, which means that the single MD group could well contain a certain number of misclassified cases, obscuring possible differences in 'type'. Whereas obviously, therefore, no conclusive evidence was obtained regarding the differences between the underlying 'true type', assuming such a thing exists, our analyses did show that having experienced more than one episode was not associated with more disability than experiencing a first - perhaps sole - episode.

This lack of difference in degree of disability can be explained in several ways. Firstly, it may be due to a response shift on the SF-36. Individuals with recurrent MD may adapt to their situation, causing them to make positive adjustments to their assessment of disability as compared to the first episode. A second factor could be differences in treatment-seeking behaviour. If individuals with recurrent MD are more often successfully treated, this can decrease their disability, diminishing differences with single MD. Thirdly, frequent episodes may well still be associated with more disability. Our study was restricted to examining differences between a single episode and recurrent episodes of MD, rather than by frequency. Finally, recurrent MD could still be associated with more disability in a clinical setting, where this idea was formed. Differences between a community setting, which was studied here, and a clinical setting may explain the outcome. Further study is needed to determine which of these explanations plays a role. Despite similar levels of disability, other parameters, such as prognosis and response to therapy, may still differ with 'type'.

We did find a relation between MD 'severity' level and disability: a higher level of 'severity' was associated with more disability. This is consistent with the DSM-III-R criteria for MD 'severity'. The observation that limitations in physical functioning were not associated with 'mild' MD, but only with higher levels of 'severity' also corresponds with this definition. Not each

increase in 'severity' coincided with more disability. 'Moderate' and 'severe' MD seemed to be associated with a similar level of disability: only a difference in mental health was found. Because the items of the mental health scale are rather similar to the DSM criteria for MD and hence closely related to the CIDI diagnosis of MD, we believe it is not an appropriate measure of disability in our study.

Individuals diagnosed with 'mild' MD differed substantially from non-depressed individuals on indicators that represent psychological and social functioning. Moreover, these differences showed larger effect sizes than the differences between the successive levels of 'severity'. Apparently, 'mild' MD causes major disability, which can be exacerbated by higher levels of 'severity'. We believe for this reason that although less severe cases may well be detected in a community setting compared to a clinical setting [13], these mild cases are nonetheless still relevant. This is consistent with the literature describing considerable disability to be associated even with subthreshold levels of depression in the community [5, 6, 28, 29]. The results reported by Henderson et al. [30] are also in line with this idea. Although they report 14% of MD cases in a community survey to be free of disability, 86% of cases did have limitations in ability to work and perform role-activities. Moreover, while Henderson et al. [30] limited their definition of disability to these latter two indicators, our study showed that 'mild' cases report limitations in indicators other than these. In view of our results the sceptisisms about using fully structured interviews [31] may be seen from a different perspective. Although the SCAN, a semi-structured interview administered by clinically trained interviewers, has been reported to detect less MD cases than the CIDI [31], the additional cases detected by the CIDI may still be relevant.

Three factors may detract from our conclusions. The first concerns the representativeness of the sample, which could be questioned for two reasons. In the first place, homeless and long-term institutionalised individuals are not included in the sample, so that some very severely depressed individuals may be missed. In the second place, selective non-response may also have caused under representation of severely depressed persons. As a result of such under representation the extent of the disability in the highest 'severity' group may be underestimated. This may be countered by the argument that inspection of the non-responders showed that their psychiatric morbidity did not differ from the responders [4, 14], although only 44% of the non-responders were available for

this analysis. Furthermore, the number of patients institutionalised with MD for more than a year is a relatively small group (1620 affective psychoses (first diagnosis) in general psychiatric hospitals in the Netherlands in 1993 [32]). We therefore assume that the sample is an adequate representation of the Dutch population of depressed individuals.

A second possible drawback is the fact that our analyses were performed on a sample that had been diagnosed with MD during the past year, whereas the SF-36 is designed to assess disability over the period of the past month. The approximate 50% of individuals that are no longer suffering from MD may obscure the relation between the SF-36 scores and 'severity' or 'type'. For this reason we repeated our regression analyses using the sample of individuals diagnosed with MD during the last month. Although the disability indicators differed less significantly with 'severity', the larger regression coefficients indicate that the relation between 'severity' and disability is even stronger when the month prevalent sample was used. The less significant p-values in the analyses of variance can be explained by the low numbers of cases included. The relation with 'type' remained insignificant. In fact, our finding that recurrent MD was not associated with more disability is strengthened by the use of the year prevalent sample. It may well be that recurrent MD is associated with more residual disability instead of with more disabling episodes of MD. Using the year prevalent sample we incorporated such residual disability into the analysis, and showed that even when it is taken into account, levels of disability are similar for single and recurrent MD.

A third limitation to our study is the use of cross-sectional data. Because we were only interested in how burdensome it is for individuals to have a depressive episode within a certain time frame (i.e. a year), not in their future prospects of disability, using cross-sectional data suffices. Longitudinal data, nevertheless, may provide further insight into the difference between single and recurrent depression MD, as a subject can be more disabled not only by having a higher degree of disability, but also by having the same degree of disability for a longer period of time.

Conclusion

The DSM-III-R diagnosis of 'severity' classes provides information on the disability associated with MD. Diagnosis of MD 'type', on the other hand, did

not. Consequently, the diagnosis of 'severity' can be used to estimate the distribution of disability in the depressed population, information that is important for health policy and planning. Broken down according to disability level, only three groups of MD could be distinguished: 'mild', 'moderate to severe', and 'severe MD with psychotic features'. A simplification of the DSM 'severity' classes could be considered. The distinct difference in disability between 'mild' MD and no MD suggests that 'mild' cases are relevant. This is important information for interpreting the high prevalence figures from community surveys.

Acknowledgements

We are specifically grateful to Dr. N.N. Nagelkerke for his assistance with the statistical analyses. Furthermore, we would like to thank Prof. P.J.M. van der Maas, Prof. D. Kromhout, Dr. P.G.N. Kramers and Dr. J.J. Barendregt for their constructive comments on this paper.

References

1. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry* 1994;51(1):8-19.
2. Offord DR, Boyle MH, Campbell D, Goering P, Lin E, Wong M, et al. One-year prevalence of psychiatric disorder in Ontarians 15 to 64 years of age. *Can J Psychiatry* 1996;41(9):559-63.
3. Australian Bureau of Statistics. *Mental Health and Wellbeing: Profile of Adults, Australia*. Canberra: Australian Government Publishing Service; 1997.
4. Bijl RV, Ravelli A, van Zessen G. Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc Psychiatry Psychiatr Epidemiol* 1998;33(12):587-95.
5. Broadhead WE, Blazer DG, George LK, Tse CK. Depression, disability days, and days lost from work in a prospective epidemiologic survey. *Jama* 1990;264(19):2524-8.

6. Kessler RC, Zhao S, Blazer DG, Swartz M. Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *J Affect Disord* 1997;45(1-2):19-30.
7. Bijl RV, Ravelli A. Current and residual functional disability associated with psychopathology: findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Psychol Med* 2000;30(3):657-68.
8. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *Jama* 1989;262(7):914-9.
9. Ustun TB, Sartorius N, editors. *Mental Illness in General Health Care: an International Study*. London: John Wiley and Sons; 1995.
10. The Counselling Versus Antidepressants In Primary Care Study Group. How disabling is depression? Evidence from a primary care sample. *Br J Gen Pract* 1999;49(439):95-8.
11. Sprangers MA, de Regt EB, Andries F, van Agt HM, Bijl RV, de Boer JB, et al. Which chronic conditions are associated with better or poorer quality of life? *J Clin Epidemiol* 2000;53(9):895-907.
12. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders : DSM-III-R*. 3rd , rev. Washington, DC: American Psychiatric Association; 1987.
13. Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, et al. Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Arch Gen Psychiatry* 1998;55(2):109-15.
14. Bijl RV, van Zessen G, Ravelli A, de Rijk C, Langendoen Y. The Netherlands Mental Health Survey and Incidence Study (NEMESIS): objectives and design. *Soc Psychiatry Psychiatr Epidemiol* 1998;33(12):581-6.
15. World Health Organization. *Composite International Diagnostic Interview (CIDI), version 1.0*. Geneva; 1990.
16. Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988;45(12):1069-77.
17. Wittchen HU. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. *J Psychiatr Res* 1994;28(1):57-84.
18. World Health Organization. *International Classification of Impairments, Disabilities and Handicaps*. Geneva; 1980.
19. World Health Organization. *International Classification of Functioning, Disability and Health*. Geneva; 2001.

20. Nagi SZ. *Disability concepts revisited: implications for prevention.* In: Pope AM, Tarlov AR, editors. Disability in America: Toward a National Agenda for Prevention. Washington DC: National Academy Press; 1991. p. 309-327.
21. Verbrugge LM, Jette AM. The disablement process. *Soc Sci Med* 1994;38(1):1-14.
22. Ware JE. *SF-36 Health Survey, manual and interpretation guide.* Boston, Mass: Nimrod Press; 1993.
23. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol* 1998;51(11):1055-68.
24. Beusterien KM, Steinwald B, Ware JE, Jr. Usefulness of the SF-36 Health Survey in measuring health outcomes in the depressed elderly. *J Geriatr Psychiatry Neurol* 1996;9(1):13-21.
25. van den Berg J, van den Bos GAM. Het (meten van het) voorkomen van chronische aandoeningen 1974-1987. *Mndber Gezondheid (CBS)* 1989;8(3):4-21.
26. Cohen J. *Statistical power analysis for the behavioural sciences.* New York: Academic Press; 1977.
27. SAS Institute Inc. *SAS/STAT User's Guide.* 6 ed. Cary, NC, USA.; 1990.
28. Beekman AT, Deeg DJ, Braam AW, Smit JH, Van Tilburg W. Consequences of major and minor depression in later life: a study of disability, well-being and service utilization. *Psychol Med* 1997;27(6):1397-409.
29. Lyness JM, King DA, Cox C, Yoediono Z, Caine ED. The importance of subsyndromal depression in older primary care patients: prevalence and associated functional disability. *J Am Geriatr Soc* 1999;47(6):647-52.
30. Henderson S, Korten A, Medway J. Non-disabled cases in a national survey. *Psychol Med* 2001;31(5):769-77.
31. Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001;31(6):1001-13.
32. piggz. *patientenregister intramurale geestelijke gezondheidszorg landelijke tabellen psychiatrie.* Utrecht: piggz; 1993.

6

The burden of depression in the Netherlands: Is disability overestimated?

Michelle E. Kruijshaar, Nancy Hoeymans, Jan Spijker, Marlies E.A. Stouthard, Marie-Louise Essink-Bot. The burden of depression in the Netherlands: Is disability overestimated? *Submitted.*

Abstract

Context:

The high burden of major depression is caused by high prevalence figures in combination with a high disability weight. The available weights, however, seem to be based on clinical descriptions of functioning, rather than on information of depression-associated disability within the community.

Objective:

To investigate whether this has led to an overestimation of the burden of depression.

Design:

Disability information from a Dutch mental health survey was used to distinguish different severity-classes of major depression and determine their proportional prevalence. We obtained expert valuations for each class and calculated an overall disability weight, weighting for the proportional prevalence.

Results:

Overall and severity-specific weights were very similar to other studies, but the overall DW was 73% higher than the weights used in the Dutch Burden of Disease Study, in which proportional prevalence was based on expert opinion.

Conclusions:

This study provides no indication to suspect that the disability and burden of depression have been overestimated. The Dutch example shows the importance of tailoring the disability weights to the epidemiological data.

Introduction

The importance of mental disorders and particularly of major depression (MD) as a cause of disease burden was one of the major findings of the Global Burden of Disease Study 1990 (GBD 1990) [1]. World-wide, depression accounted for nearly 11% of the disability and 4% of the total disease burden. The measure of burden of disease used in this study, the Disability-Adjusted Life-Year (DALY), combines the number of life years lost due to premature mortality and the number of years lived with disability using a set of disease-specific disability weights (DWs). Years lived with a specific disease are weighed for the severity of the disability that is associated with the disease using DWs.

In several national burden of disease studies that followed the GBD-1990, the prominent burden of MD was replicated [2-4]. The prominence of MD reflects both the high prevalence figures found in community surveys [5-7], and the strong effects of MD on functioning and well-being [7-11].

In the original DALY methodology of the GBD-1990 DWs were determined by health experts. While the epidemiological estimates such as incidence or prevalence were hardly debated, the valuation method used to derive the DWs, the use of expert values and the universality of the DWs, have been criticised [12, 13]. For depression and other mental disorders, there are additional problems specific to the field of mental disorders that concern the DWs.

One problem is that the estimates of the burden of depression are based on prevalence data from community surveys. This requires the DWs for depression to reflect the situation in the community. However, much of the empirical information on the functional effects of MD relates to clinical cases. Population surveys, on the other hand, may include milder cases of depression than those found in clinical settings [14], and DWs may as a consequence be overestimated. Having a low mortality component [1-4], the burden of MD relies heavily on the DWs and prevalence estimates, and the effect of overestimating the DW will be large.

The aim of this study was to investigate whether accurate tailoring of the DWs to the prevalence data decreases the estimated burden of depression. To study this we needed information on the distribution of disability among the depressed cases in the community. Previously, we examined this using data from

a Dutch general population sample [15]. Disability was defined as limitations in the physical, psychological and social domains of functioning. Three clusters of "severity" could be distinguished: "mild", "moderate to severe" and "MD with psychotic features" [15]. We derived DWs for these three severity-classes, compared the DWs with other studies, and calculated the burden of depression using these tailored DWs.

Methods

Derivation of Disability Weights

– *Disease selection, staging and description*

We presented five disorders for valuation: MD, obsessive compulsive disorder, cancer of the oesophagus, prostate cancer and vision disorders. We included disorders other than MD to prevent accentuating MD. Each disease was subdivided into assumed homogeneous stages regarding disability, treatment and prognosis. In total 18 disease stages were valued: the three severity-classes of MD (see introduction), 3 stages for oesophageal cancer, 2 for obsessive compulsive disorder, and 5 for the other two disorders.

For each disease stage, a lay text and a standardised generic description of the functional health status was provided. An example is shown in Figure 6.1. We used the EuroQol 6D5L to describe the generic health status [16, 17]. This system describes six dimensions (6D) of health: mobility, self-care, usual activities, pain/ discomfort, anxiety/ depression and cognition, in five levels (5L) of severity of the problems. The first, third and fifth levels are identical to the EuroQol 6D3L; two intermediate levels are inserted.

For MD we based the lay text on the DSM-III-R criteria for MD and "severity" of MD. EuroQol descriptions were based on the reported disability in the Netherlands Mental Health Survey and Incidence Study (NEMESIS) [8, 15]. From this study we used as indicators of disability the eight scales of the Short-Form-36 Health Survey (SF-36) [18], the number of days spent in bed due to psychiatric, drug or alcohol related problems, and the number of days not able to work due to these problems. We used a formal algorithm (available from the authors on request) to map these disability data on the EuroQoL 6D5L classification.

Figure 6.1. **Example of a disease stage description.**

Depression

Depression is divided into the following stages:

1. mild,
2. moderate to severe,
3. severe with psychotic features.

We now ask you to value:

Patients with moderate to severe depression

These people experience one or more depressive episodes within a year. During these periods they go through permanent feelings of sadness or emptiness and a permanent loss of interest or pleasure in nearly all activities. They have problems eating and/ or sleeping and feel worthless or guilty. They may have thoughts of death.

In a year in which one or more episodes are experienced their condition is:

	No problems in walking about
	No problems with self care
●●	Some problems with performing usual activities *(e.g. work, study, housework, family or leisure activities)*
●●	Moderate pain or discomfort *(fatigue)*
●●	Moderately anxious or depressed
●●	Some cognitive impairments *(memory, concentration, desorganisation, IQ-level)*

The descriptions for obsessive compulsive disorder and oesophageal cancer were adopted from a previous study: the Dutch Disability Weights Study (DDW) [19]. We recoded their EuroQol 6D3L descriptions into 6D5L. The descriptions and resulting values for prostate cancer and vision disorders will not be discussed here.

- *Valuation procedure and respondents*

The valuation procedure was largely copied from the DDW Study [19]. In brief, we employed medical doctors with sufficient knowledge of the consequences of a broad range of diseases. A convenience sample of 75 doctors was contacted, 26 of whom had previously participated in similar studies [19, 20].

A written postal questionnaire was used. We replicated the DDW's "interpolation" procedure [19], in which subjects were asked to interpolate disease stages on a disability scale. The scale ranges between 0 (worst imaginable health state) and 100 (best) and was formally calibrated with PTO-derived weights for 16 conditions. On the scale we replaced the conditions "mild MD" and "severe vision disorder" by disorders with comparable DWs ("mild to moderate panic disorder" and "grade 3-4 arthritis").

The duration of a disease stage to be valued was defined universally as one year. For MD, with its episodic character, we asked the respondents to value: "a year in which one or more depressive episodes took place". The general description then showed the average health status for individuals who had an episode of MD in the last year.

- *Analyses of the interpolation data*

For each disease stage, we calculated the DW from: $1-mean\_value/100$. We examined the reliability of the valuations by 1) checking the compliance with the imposed ordering of the stages of mental disorders ("mild", "moderate", "severe"); 2) inspecting the Spearman rank correlation and inter-rater reliability among respondents; and 3) estimating the proportion of total variance that was attributable to the disease stages applying generalisability theory (G-study) [21, 22].

We also studied possible associations of age, sex, current profession (GP, psychiatrist, researcher, other), and having medical experience (less versus more than 1 year) with the valuations in a regression analysis. All analyses were performed in SAS version 6.12 [23].

- *Comparison to the Dutch Disability Weights Study (DDW)*

We compared our DWs for MD with those from the DDW. Because respondents of the two studies were not independently drawn, standard statistical testing does not apply. Nevertheless to get an impression of the differences we compared the 95% confidence intervals. We also compared the weights for obsessive

compulsive disorder and oesophageal cancer, to estimate the level of retest-variability.

Burden of Disease

– *An overall weight for MD*

An "overall" average DW for MD was calculated from the severity-specific DWs, weighting for their proportional prevalence. Prevalence data for the three severity classes were obtained from the NEMESIS study. The residual prevalence of cases with "unspecified severity" in this study was proportionally distributed over the classes, excluding "MD with psychotic features", as we assumed these cases were unlikely to be missed.

- *Calculation*

We calculated the burden of MD in the Netherlands in 1996 from the number of years lived with disability (YLD) only. We ignored mortality, because mortality from depression is typically low, and was also not included in the Dutch National Burden of Disease Calculation for 1994 [3] (suicide was measured separately in this study).

We estimated YLDs for MD by multiplying 1-month prevalence and the overall DW with 1996 population data from the Netherlands Bureau of Statistics. Month prevalence figures were used as estimates of point prevalence. For adults aged 18-64 years we obtained 1996 prevalence data from the NEMESIS study [6]. We supplemented these for 13-17 year olds using 1993 figures [24] and for those aged over 65 using 1992/3 prevalence data [25]. The 6-month prevalence measured for the adolescent group was converted to 1-month prevalence using the ratio of 1 to 6-month prevalence from the NEMESIS study.

- *Comparison with previous estimates*

We compared the overall DW with weights from four studies: the Dutch National Burden of Disease Calculation for 1994 [3], the GBD-1990 [1], the Australian Burden of Disease Study [2], and a study by Andrews et al. [26]. Both the Dutch and Australian Burden of Disease Studies used severity-specific weights from the DDW to calculate an overall DW, but applying different proportional distributions of the severity-classes. We compared the number of DALYs lost to MD in the Netherlands in 1996 calculated using our tailored DW and the one from the Dutch Burden of Disease Study.

Results

Derivation of Disability Weights

– *Description of MD classes*

The formal and lay descriptions of the three severity-classes of Major Depression (MD) are given in Table 6.1.

- *Respondents*

In total 49 medical experts participated (24 men and 25 women), a response rate of 64%. The mean age was 46.6 years (SD 8.8), the mean number of years of medical experience 12.2. Of the respondents, 53% was involved in direct patient care (14 GPs, 5 psychiatrists and 7 others), 35 % worked in medical research and 12% had other health-related professions or was retired.

- *Analyses*

The top part of Table 6.2 shows the DWs and their 95% confidence intervals (95%-CI) for the three classes of depression. The rank-order implied by the severity-specific classes of the psychiatric disorders (mild/ moderate/ severe/ severe with psychotic features), was complied with, except by one respondent who valued "MD with psychotic features" equal to "moderate to severe". Respondents largely agreed with each other on the ranking of the 18 disease stages: the mean Spearman correlation coefficient was 0.83. Also, individual valuations correlated well with the rest: Pearson correlation coefficients were larger than 0.8 except for one.

In the variance component analysis, 76% of total variance was explained by the disease stages. Respondents contributed another 6%, while a residual 18% remained unexplained. Regression analyses showed that age, sex, current profession, and not having medical experience could not significantly predict the scores of the 18 disease stages.

- *Comparison to the Dutch Disability Weights Study (DDW)*

In the top part of Table 6.2 we also provide the DWs for MD obtained in the DDW [19]. Our severity-class descriptions and the number of classes differed from the DDW. Our EuroQol descriptions were generally less severe than in the DDW. Our DWs fell within the range of the 95%-CI of the DDW values, except for "moderate to severe" MD, where it is in between the 95%-CIs of the two separately valued classes.

Table 6.1. **Descriptions of the three severity-classes (stages) of major depression (MD).**

MD class "mild"	(EuroQol 6D5Ldescription:112222)
These people experience one or more depressive episodes within a year. During these periods they go through permanent feelings of sadness or emptiness or a permanent loss of interest or pleasure in nearly all activities. They can have problems eating or sleeping and can feel worthless or guilty. They may have thoughts of death.	
MD class "moderate" to "severe"	(EuroQol 6D5Ldescription: 113333)
These people experience one or more depressive episodes within a year. During these periods they go through permanent feelings of sadness or emptiness and a permanent loss of interest or pleasure in nearly all activities. They have problems eating and/ or sleeping and feel worthless or guilty. They may have thoughts of death.	
MD class "severe with psychotic features"	(EuroQol 6D5Ldescription: 214444)
These people experience one or more depressive episodes within a year. During these periods they go through permanent feelings of sadness or emptiness and a permanent loss of interest or pleasure in nearly all activities. Furthermore, they experience delusions and hallucinations. They have problems eating and sleeping and feel worthless or guilty. They have thoughts of death.	

*1. no problems, 2. a few problems, 3. some/ moderate problems, 4. severe problems, 5. very severe problems/ inability. The six digits correspond to the six dimensions of the EuroQol 6D: mobility, self-care, usual activities, pain/ discomfort, anxiety/ depression, and cognition.

Also shown in Table 6.2 are the DWs for obsessive-compulsive disorder and cancer of the oesophagus. Our descriptions for these disorders were taken from the DDW. Re-valuation of these stages resulted in new average values that fell within the 95%-CIs of the DDW, except for severe obsessive-compulsive disorder.

Table 6.2. **Disability Weights obtained in the present study and comparison to the Dutch Disability Weights Study.**

Disease stage	This study			DDW		
	EuroQol	DW	95% CI	EuroQol	DW	95% CI
Major Depression:						
- "Mild"	112222	0.19	0.16-0.22	113131	0.14	0.09-0.19
- "Moderate" to "severe"	113333	0.51	0.46-0.55			
- Moderate				133133	0.35	0.27-0.42
- Severe				335353	0.76	0.56-0.97
- "With psychotic features"	214444	0.84	0.80-0.88	335355	0.83	0.75-0.92
Cancer of the oesophagus						
- Stage of diagnosis and primary therapy (pt)	112441	0.52	0.48-0.57	112441*	0.53	0.36-0.70
- State after intentionally curative pt	113331	0.42	0.37-0.46	113331	0.38	0.25-0.51
- Irradically removed / disseminated carcinoma	114451	0.82	0.79-0.84	114451*	0.73	0.61-0.86
Obsessive-compulsive disorder						
- Mild to moderate	113133	0.30	0.26-0.33	113133	0.24	0.17-0.32
- Severe	133155	0.76	0.71-0.82	133155	0.56	0.38-0.74

DW: Disability Weight; DDW: Dutch Disability Weights Study [19].
*For these stages the DDW descriptions gave two descriptions in EuroQol 6D3L, each with 50% possibility. We recoded these into EuroQol 6D5L levels by averaging the two descriptions, thereby using the two additional levels (level 2 and 4) of this system.

For MD the differences with the DDW weights do not appear to be larger than for the identically described disorders. Most new weights fell within the DDW's 95%-CIs, and the absolute differences with the DDW for MD (0.01 and 0.05) were smaller than for the other diseases (between 0.01 to 0.20). For all three disorders, our 95% confidence intervals were smaller than in the DDW.

Burden of disease

– *Overall DW for depression*

Table 6.3 shows how we calculated an average disability weight for MD. Weighting for the prevalence distribution of depressive cases across the severity-classes we calculated an overall DW of 0.46.

- *Burden of Disease*

We estimated that a total of 151,137 disability adjusted years were lost to MD in 1996 in the Netherlands; 9.7 per 1000 persons (total population of 15,5 million). Per 1000 men 6.8 DALYs were lost to MD and per 1000 women 12.6.

- *Comparison to previous estimates*

In Table 6.4 we compare the overall weight for MD to other studies. Our estimate is similar to the GBD weight and close to the weights from two Australian studies. However, our overall weight for MD is 73% higher than the one used in the 1994 Dutch National Burden of Disease Calculation. In that study the severity-specific weights from the DDW were combined to an overall DW for MD using expert estimations of the proportional prevalence of the severity-classes: 60% for "mild", 30% "moderate", 9% "severe" and 1% for "severe with psychotic features. Although our severity-specific DWs did not differ largely from the DDW, the NEMESIS data on the proportional distribution of the severity-classes that we used to combine them (see Table 6.3) were very different.

Using the DW from the National Burden of Disease Study [3] the estimated number of DALYs lost to MD (excluding dysthymia) is 5.6 per 1000 persons, 42% lower than the number calculated using the tailored DW.

Table 6.3 **Calculation of the average disability weight for Major Depression.**

Severity-class of depression	% of prevalence*	DW
"Mild"	26.7**	0.19
"Moderate" to "severe"	61.9**	0.51
"Severe with psychotic features"	11.4	0.84
Total MD	100	0.46

DW: Disability Weight, MD: major depression.
*Prevalence data obtained from the Netherlands Mental Health Study and Incidence Survey (NEMESIS) [6, 15]. **Including MD of "unspecified severity".

Discussion

We derived DWs for MD that are tailored to the level of MD-associated disability observed in a community sample in the Netherlands. For the separate severity-classes of MD our DWs did not deviate largely from the Dutch Disability Weights Study (DDW) [19]; even though our health status descriptions were more severe. Our overall weight for MD was similar to those used in several burden of disease studies [1, 2, 26]. Due to differences in the estimated proportional prevalence of the severity-classes, nevertheless, our overall DW was 73% higher than the one used in the 1994 Dutch National Burden of Disease Calculation [3]. As a result, we estimated that per 1000 persons 9.7 years lived with disability (YLDs) were lost to MD in the Netherlands in 1996.

We derived DWs for three severity-classes of depression providing descriptions of the health states that were based on information obtained in a Dutch community survey [8, 15]. Because our health status descriptions in EuroQol (based on self-reported disability in the survey) were on average somewhat less severe than the ones provided in the DDW (based on case-definition and expert opinion), we expected our DWs to be lower (i.e. indicating less disability) than the DDW weights. Nevertheless, the differences between the two studies did not appear significant. Similar weights for depression were also found in an Australian study [27], in which the health status was provided as Short Form-12 (SF-12) descriptions, that were obtained from the Australian

Table 6.4. **Comparison of the overall disability weights for depression from different studies, and how they were derived.**

Study, country	DW (method)	Severity-classes	Distribution across classes	DW
This study, the Netherlands	New (ip-PTO)	3 severity classes: "mild", "moderate to severe", "severe with psychotic features".	Dutch survey data*	0.459
GBD-1990, EME [1]	GBD (PTO)	2 classes: treated versus untreated.	expert estimation	0.469
Dutch National Burden of Disease Calculation [3]	DDW (PTO & ip)	4 Severity classes: "mild", "moderate", "severe", "severe with psychotic features".	expert estimation	0.266
Australian Burden of Disease Study [2]	DDW (see DDW)	3 Severity classes: "mild", "moderate", "severe".	SF-12 cut-off scores in Australian survey**	0.41 (m) 0.37 (f)
Andrews et al., Australia [26, 27]	New (PTO)	3 Severity classes: "mild episode", "moderate episode", "severe episode".	SF-12 cut-off scores in Australian survey**	0.417

DW: Disability Weight; PTO: person trade-off; ip-PTO: interpolation on a PTO calibrated disability scale; GBD-1990: Global Burden of Disease Study 1990; EME: established market economies; DDW: Dutch Disability Weights Study; m: male; f: female.

*The Dutch Survey data were derived from The Netherlands Mental Health Study and Incidence Survey [6, 15].

**The Australian Survey data were collected in The Australian National Mental Health and Wellbeing Survey [7].

Mental Health Survey [7]. Apparently, the health status description has only little effect on the valuation, and the disease-stage label is much more important.

Our overall DW was similar to those used in the GBD and in two Australian studies. Their correspondence suggests that there is no reason to suspect that DWs for MD, which are not tailored to the community situation, are overestimated. Therefore, the high burden of depression as estimated in the GBD and several national studies does not appear to be overrated due to an overvaluation of the DWs. On the contrary, in the Dutch 1994 National Burden of Disease Calculation the burden of depression seems to be too low as a result of underestimating the overall DW. In the latter study the proportional distribution of the severity-classes was based on expert opinion. The proportion of "mild" cases estimated by experts was much higher than the NEMESIS survey data suggest. This altered the overall DW severely (73%), showing us the importance of the quantitative epidemiological information in burden of disease calculations. The overall DW for depression was more sensitive to the distribution of disability across the depressed population, than to differences in health status descriptions.

Three remarks may be put forward to detract from our conclusions. Firstly, our generic health status descriptions were based on disability reported by people who were depressed in the past year. Thus, not all cases included were currently depressed, and disability estimated in this way is less severe than when only current cases had been included. However, there were three reasons why we did not use current cases only. Firstly, it enabled us to include residual disability, secondly to increase the number of observations and, thirdly, to provide a description of the health status in a year. The last reason is important, as the time-frame for which health states were valued was, as in most valuation studies, one year. The fact that disability is possibly underestimated in this way only strengthens our conclusion that DWs for MD were not overestimated in previous studies.

Secondly, the use of these descriptions of health status over the past year has further implications. Strictly speaking, DWs for these "annual-profiles" cannot be compared to the ones from the DDW that were derived for a whole year living with the health status at the actual moment of a depressive episode. Furthermore, to calculate the burden of MD, the annual-profile DWs should be multiplied with an annual prevalence. However, we compared with the DDW and used estimates of point-prevalence. Our rationale for this was that

respondents seem to respond mainly, or entirely, to the provided disease-stage label, not paying much attention to the generic health status descriptions and annual profile.

Thirdly, if the respondents mainly valued their idea of the average "mild", or "moderate" case of depression, it is not possible to tailor the severity-specific DWs to the community setting. Should the robustness of the DWs lead us to believe that the burden of depression is not overestimated, or are expert opinions insensitive to the subtle differences in the descriptions provided? Nevertheless, irrespective of this stand the tailoring of the overall DW using the proportional distribution of the classes, and the importance of the epidemiological data.

Conclusions

Our study provides no indication that previously estimated DWs were overestimates because they were not tailored to the community setting. Our tailored DWs were similar to most other studies, including the GBD and do not decrease the estimated burden of depression. These results strengthen the validity of the high estimates of the burden of depression. This study furthermore points out the importance of obtaining sound epidemiological data in burden of disease studies.

Acknowledgements

This study was sponsored by the Netherlands Institute of Health Sciences. The Erasmus University Trustfund financially supported a presentation at the 15th REVES (International Network on Health Expectancy and the Disability Process) Conference, May 5-7 2003, Guadalajara, Mexico.

References

1. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge: Harvard University Press; 1996.

2. Mathers CD, Vos ET, Stevenson CE, Begg SJ. The Australian Burden of Disease Study: measuring the loss of health from diseases, injuries and risk factors. *Med J Aust* 2000;172(12):592-6.
3. Melse JM, Essink-Bot ML, Kramers PG, Hoeymans N. A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *Am J Public Health* 2000;90(8):1241-7.
4. Michaud CM, Murray CJ, Bloom BR. Burden of disease--implications for future research. *JAMA* 2001;285(5):535-9.
5. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry* 1994;51(1):8-19.
6. Bijl RV, Ravelli A, van Zessen G. Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc Psychiatry Psychiatr Epidemiol* 1998;33(12):587-95.
7. Andrews G, Henderson S, Hall W. Prevalence, comorbidity, disability and service utilisation. Overview of the Australian National Mental Health Survey. *Br J Psychiatry* 2001;178:145-53.
8. Bijl RV, Ravelli A. Current and residual functional disability associated with psychopathology: findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Psychol Med* 2000;30(3):657-68.
9. Broadhead WE, Blazer DG, George LK, Tse CK. Depression, disability days, and days lost from work in a prospective epidemiologic survey. *JAMA* 1990;264(19):2524-8.
10. Kessler RC, Zhao S, Blazer DG, Swartz M. Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *J Affect Disord* 1997;45(1-2):19-30.
11. Sprangers MA, de Regt EB, Andries F, van Agt HM, Bijl RV, de Boer JB, et al. Which chronic conditions are associated with better or poorer quality of life? *J Clin Epidemiol* 2000;53(9):895-907.
12. De Wit GA, Busschbach JJ, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? *Health Econ* 2000;9(2):109-26.
13. Ustun TB, Rehm J, Chatterji S, Saxena S, Trotter R, Room R, et al. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* 1999;354(9173):111-5.
14. Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, et al. Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Arch Gen Psychiatry* 1998;55(2):109-15.

15. Kruijshaar ME, Hoeymans N, Bijl RV, Spijker J, Essink-Bot ML. Levels of disability in Major Depression. Findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *J Affect Disord* 2003;77(1):53-64.
16. Janssen MF, Bonsel GJ. Should the Euroqol descriptive system be extended from three to five levels? A methodological strategy with an empirical pilot. In: *International Society for Pharmacoeconomics and Outcome Research*; 2002; Rotterdam; 2002.
17. Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol* 1999;52(4):293-301.
18. Ware JE. *SF-36 Health Suryvey, manual and interpretation guide*. Boston, Mass: Nimrod Press; 1993.
19. Stouthard MEA, Essink-Bot ML, Bonsel GJ, on behalf of DDW Group. Disability weights for diseases in the Netherlands. *Eur J Public Health* 2000;10:24-30.
20. Essink-Bot ML, Pereira J, Packer C, Schwarzinger M, Burstrom K. Cross-national comparability of burden of disease estimates: the European Disability Weights Project. *Bull. World Health Organ.* 2002;80(8):644-52.
21. Krabbe PF, Essink-Bot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. *Soc Sci Med* 1997;45(11):1641-52.
22. Streiner DL, Norman GR. *Generalizability theory*. In: Streiner DL, Norman GR, editors. Health Measurement Scales. A Practical Guide to Their Development and Use. second edition ed. New York: Oxford University Press Inc.; 1995. p. 128-143.
23. SAS Institute Inc. *SAS/STAT User's Guide*. 6 ed. Cary, NC, USA: SAS Institute Inc.; 1990.
24. Verhulst FC, van der Ende J, Ferdinand RF, Kasius MC. The prevalence of DSM-III-R diagnoses in a national sample of Dutch adolescents. *Arch Gen Psychiatry* 1997;54(4):329-36.
25. Beekman AT, Deeg DJ, van Tilburg T, Smit JH, Hooijer C, van Tilburg W. Major and minor depression in later life: a study of prevalence and risk factors. *J Affect Disord* 1995;36(1-2):65-75.
26. Andrews G, Sanderson K, Corry J, Lapsley HM. Using epidemiological data to model efficiency in reducing the burden of depression. *J Ment Health Policy Econ* 2000;3(4):175-186.
27. Sanderson K, Andrews G. Mental disorders and burden of disease: how was disability estimated and is it valid? *Aust N Z J Psychiatry* 2001;35(5):668-76.

7

The breast cancer related burden of morbidity and mortality in six European countries: The European Disability Weights projects

Michelle E. Kruijshaar, Jan J. Barendregt, and the European Disability Weights Group, The breast cancer related burden of morbidity and mortality in six European countries: The European Disability Weights projects. *The European Journal of Public Health, in press.*

Abstract

Background:

The burden of breast cancer expressed in Disability Adjusted Life Years (DALYs) was compared for six European countries and its sensitivity to different sources of variation examined.

Methods:

DALYs were calculated using country-specific epidemiological data and European Disability Weights. Epidemiological data for 1996 were obtained for Denmark, England and Wales, France, the Netherlands, Spain and Sweden. Disability weights were empirically derived.

Results:

Denmark and the Netherlands lost the largest number of DALYs (approximately 1100 DALYs per 100,000 women). They were followed by England (87% of the Danish burden), France (72%), Sweden (68%) and Spain (67%). 70 to 80% of the burden was caused by mortality. Cross-national variation in disease epidemiology was the largest source of variation in the burden of breast cancer. Variation in disability weights and uncertainty in epidemiological data had smaller effects.

Conclusion:

To compare the burden of breast cancer and most other types of cancer mortality rates provide sufficient information.

Introduction

Breast cancer is an important health problem among women in developed countries. In the Global Burden of Disease Study 1990 (GBD 1990) [1] it was among the ten leading causes of burden of disease for women in the established market economies. The summary measure used in this study, the Disability Adjusted Life-Year (DALY), combines the burden caused by premature mortality and non-fatal health outcomes into one figure. In the GBD 1990 study the DALY is calculated by summing the number of years of life lost (YLL) and the number of years lived with disability (YLD) attributable to a specific disease. The estimations are based on disease-specific epidemiological frequency data, expert derived disability weights (DW) and demographic data. The DWs weigh the years lived with a specific disease by the severity of the disability associated with it.

The DALY concept is criticized on many points. The use of universal, expert derived DWs has raised much debate regarding the reliability of the DALY, [2-5] while the reliability and availability of the epidemiological data have commonly been taken for granted. Recently, the stability of DWs was investigated in the European DWs study (EDW) [6]. Prior to this study, no country-specific DWs were available except for the Netherlands [7], and national burden of disease studies used either the GBD 1990 DWs, the Dutch weights or a combination of both [8-10].

Another point of concern is the potentially low sensitivity of the DALY to reflect true variation in the light of the error variation. True variation is caused by true differences in epidemiology, DWs or demography over time or place, whereas causes of error variation are sampling and measurement error, and incomparability in both epidemiological data and DWs. Between European countries we may expect true variation in the burden of breast cancer for three reasons. First, as a result of differences in exposure to risk factors (e.g. age at first pregnancy [11]), variation in screening practices and therapeutic management, the epidemiological frequency data differ. Secondly, due to cross-national variation in preferences for health states, DWs of breast cancer may differ. Thirdly, the countries differ in population size and structure.

The objective of this paper is twofold:

1. To compare the burden of breast cancer in six European countries.
2. To assess the relative impact of the three sources of variation on the estimated DALYs for breast cancer, and, specifically, of the DWs.

Methods

General approach

We compared the burden of female breast cancer between Denmark (Dk), England and Wales (E&W), France (Fr), the Netherlands (Nl), Spain (Sp) and Sweden (Sw) for the year 1996. This was the most recent year for which epidemiological frequency data were available for most countries. We computed YLLs using a standard expected years of life lost. [1, 12] YLDs were obtained applying an incidence perspective; meaning incidence and duration were multiplied. DALYs were the summation of the age-specific YLLs and YLDs.

In the EDW project DWs were derived for five phases of breast cancer: (1) diagnosis and primary breast conserving therapy; (2) diagnosis and mastectomy; (3) clinically disease-free without permanent sequelae; (4) clinically disease-free with major permanent sequelae; and (5) recurrent or disseminated disease (for a description and DWs see Addendum I). We calculated YLDs for each phase, necessitating the estimation of incidence and duration for each phase.

We calculated the burden of breast cancer from:

1) epidemiological frequency data on incidence and mortality,
2) DWs from the EDW project,
3) demographic data,
4) additional data to estimate incidence and duration of the five phases, and
5) a model to combine these data into a burden of disease calculation.

The baseline calculations used country-specific epidemiological data, European DWs and a standard European population (I). Additional calculations were done with alternative epidemiological frequency data (II), country-specific DWs instead of European DWs (III), and country-specific demographic data (IV), to explore the effects of uncertainty in the epidemiological estimates (II), of

introducing cross-national variation in DWs (III) and in demography (IV), respectively.

Incidence and mortality data

National incidence and cause-specific mortality data of female breast cancer (ICD-10: C-50) were collected for each country per 5 year age group up to 85+. Incidence numbers refer to the number of tumours and not women. Because women may experience a second primary tumour in their breast, incidence was adjusted for double counting of persons using estimates of the percentage of secondary breast cancer primaries (SPs). A SP was defined as a new tumour in the breast of a woman who had been diagnosed with a malignancy in the breast before. Information on the data and their sources can be found in Addendum II.

To study the sensitivity of the DALY to uncertainty in the epidemiological information, we increased the Dutch mortality rates by 5% in an additional calculation, and used regional registrations of incidence as an alternative to hospital discharge data for Spain.

Disability weights

In the EDW project, disability weights were derived from health professionals and non-health professionals in five countries (232 participants in total) using the time trade-off method and a visual analogue scale. For a description of the methods see Essink-Bot et al. [6] Empirical DWs were not derived in Denmark. For the remaining five countries country-specific weights were derived for the five phases of breast cancer (see Addendum I). European weights were calculated based on the pooled data of the five countries.

Demographic data

The European standard population, obtained from the World Health Statistics Annual 1990 [13], has a total population size of 100,000. We collected population numbers and total mortality data per 1 year age group for each country. Information on these data and the data sources are given in Addendum II.

Additional data

To estimate the incidence of each phase of breast cancer data were collected on the proportion of incident cases diagnosed as metastasized (M1) and on the proportion of women undergoing total mastectomy. We used one estimate for all age groups. Data sources are provided in Addendum II. For England, cancer

epidemiologists suspected the data on the percentage M1 to be too low, and we used a maximum estimate they provided in an alternative calculation.

Phase-specific duration was in part estimated from a study by the Munich cancer registry [14]. The diagnosis and therapy phase was assigned a duration of one year. The duration of the remaining phases was back calculated using the total disease duration (see below). The disease-free phase was allowed to last a maximum of five years. For more information see Addendum III.

Calculation of the burden of breast cancer

Incidence and mortality rates were obtained by dividing age-specific numbers by the midyear population. Incidence rates for France were extrapolated from 75+ to 85+ using cubic-spline methodology [15] and life-table derived mean ages of 87 and 93 for age groups 75+ and 85+, respectively. The proportion of SPs was subtracted from the incidence rates.

Next, the average total disease duration by age was estimated using the DisMod II software [16, 17], interpolating the rates to 1 year age groups (a good approximation of the transition hazards) using cubic-spline methodology. To obtain incidence hazards among susceptibles we divided incidence by prevalence, calculated using the model described by Barendregt et al. [18]. Both models were specified for a situation without remission and assumed that no time trends in the transition hazards exist.

Subsequently, incidence was distributed over the five phases of breast cancer using a model that describes different pathways that incident patients may pass through over time (see Addendum III). Briefly, the total proportion of patients dying from breast cancer (non-survivors) was first estimated by dividing the total number of breast cancer deaths and incident cases, after standardising them to the European standard population. Second, assuming that death from breast cancer is always preceded by metastasis or local recurrence, we calculated the proportion of non-survivors with and without metastasis at diagnosis using the percentage M1. Third, survivors and M0 non-survivors were distributed into those with and without permanent sequelae, assuming that the proportion of cases with permanent sequelae can be approximated by the proportion of women undergoing mastectomy treatment. In each pathway several phases may be passed through over time.

YLDs were calculated for each phase in each pathway by multiplying pathway-specific incidence, phase-specific duration and DW, with population

numbers in each age group. Summation resulted in estimates of total YLDs by age.

YLLs were calculated using standard expected years of life lost derived from the standard West Level 26 life-table and estimates of mean age at death per 5 year age group (Drs T. Vos, department of human services). Standard years of life lost were multiplied by mortality rates and population numbers to yield YLLs. Summation of YLDs and YLLs produced the country specific DALYs.

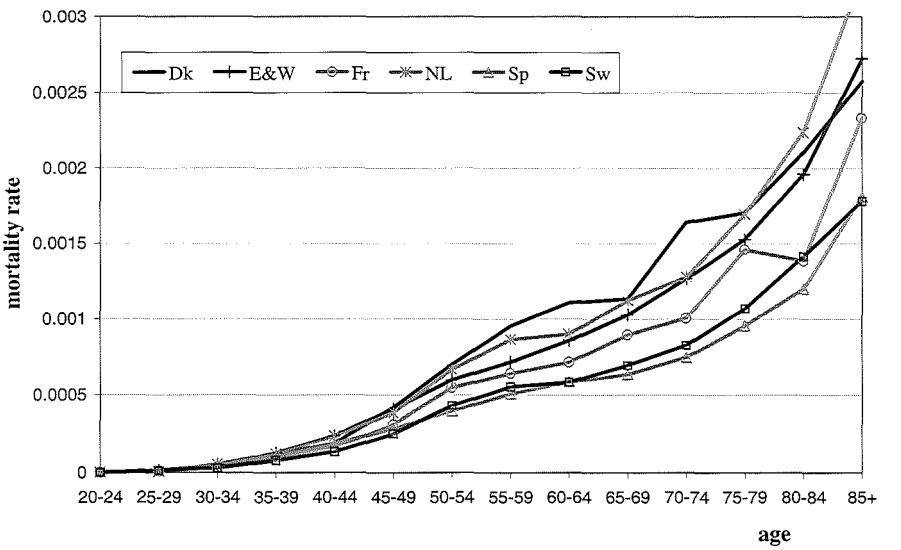
Results

Country-specific epidemiological data

Figures 7.1 and 7.2 show the observed mortality and incidence rates for each country by age. Mortality increased with age. Highest rates were reported for Denmark and the Netherlands, followed by England and Wales, while the lowest were seen in Sweden and Spain.

The incidence rates in figure 7.2 show an initially exponential increase with age, which changes around age 50 for most countries (Clemmensen's hook).

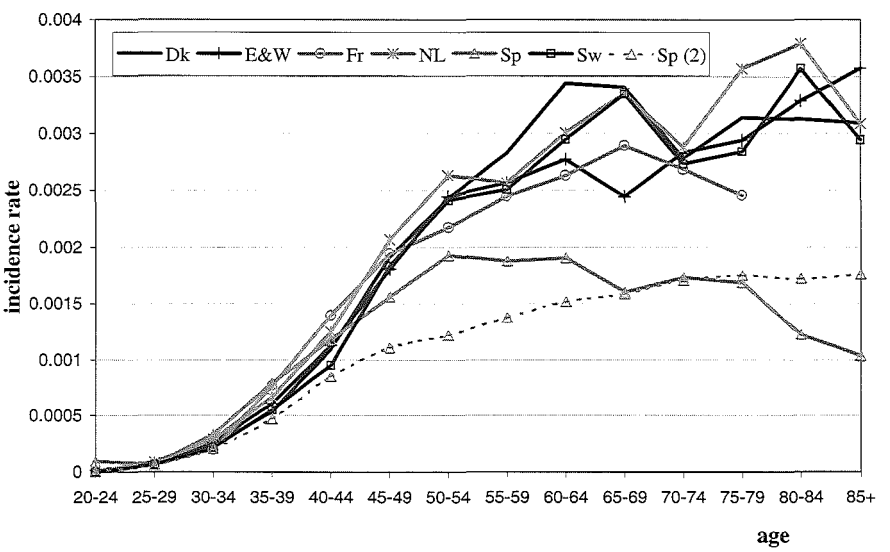
Figure 7.1. **Breast cancer mortality rates by age for women in six European countries.**



For some countries, another interruption of the exponential increase can be seen around age 65, while at the highest age groups incidence starts to decline except in England and Wales. The Spanish incidence rates were lower than in other countries from age 40-45 onwards. The 1996 Spanish hospital discharge data decreased after age 50, while the averaged regional estimates used as alternative data (dotted line) continued to increase up to age 75.

The proportion of M1 cases ranged between 2% (Sweden) and 6% (the Netherlands). For England and Wales cancer epidemiologists expected the registered 2.5% to be an underestimate, and estimated it to be maximally 10%. This maximum was used as an alternative. Therapy practice (mastectomy versus breast-conserving therapy) differed somewhat between countries. In England and France, respectively 32% and 35% of women underwent total mastectomy. In other countries the percentage ranged between 43% (Sweden) and 59% (Spain). Because cancer registration experts suspected no large deviations, Dutch regional data on the percentage SP were used for all countries except Spain. The percentage increased with age to nearly 10% in age group 70-75. In Spain, experts estimated SP as close to zero.

Figure 7.2. **Breast cancer incidence rates by age for women in six European countries.**



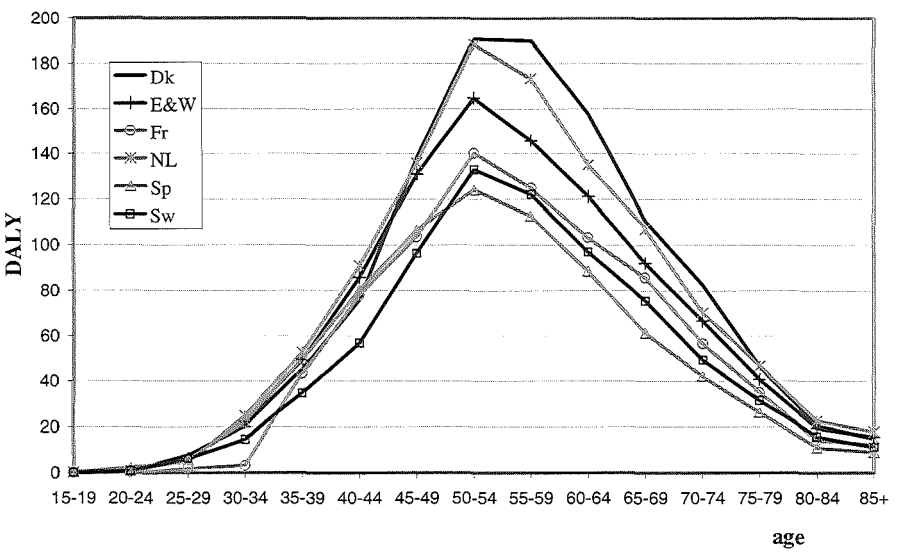
The burden of breast cancer I: epidemiological frequency data as the only source of variation

The number of DALYs calculated from country-specific epidemiological data, European DWs and a standard European population are shown by age for each country in figure 7.3. The total number of DALYs, YLDs and YLLs are given by country in table 7.1 and figure 7.4. Denmark and the Netherlands lost the highest number of years to breast cancer, followed by England and Wales (87% of the Danish burden), France (72%), Sweden, (68%) and Spain (67%). DALYs lost to breast cancer peak at around age 50. Table 7.1 also provides the percentage of DALYs attributable to mortality: in Sweden the lowest proportion of DALYs was attributable to YLLs: 70%.

The burden of breast cancer II: replacing some uncertain epidemiological data by alternative data

For England and Wales, the Netherlands and Spain the burden was calculated assuming alternative epidemiological data. The results and changes with the original calculations are depicted in table 7.2 (top). The largest deviations from the baseline estimates were observed for Spain where the hospital discharge data

Figure 7.3. **Population standardized DALYs for breast cancer by age in six European countries.**

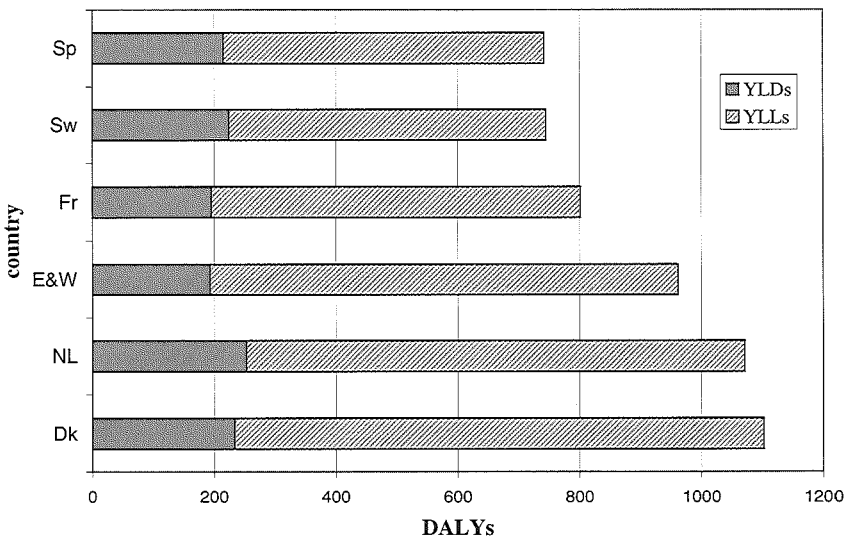


were replaced by regional incidence rates. As a result, YLLs decreased by almost 30% and the number of DALYs by 8.5%. For the Netherlands, the mortality rates were increased by 5%. This increased the number of DALYs by 6.3%. Varying the proportion of patients metastasised at diagnosis (M1) in England and Wales increased the number of YLD by only 0.2%. Using the higher mortality rates for the Netherlands caused the burden of breast cancer in the Netherlands to become slightly larger than in Denmark.

The burden of breast cancer III: cross-national differences in epidemiological data and DWs

Table 7.2 (middle) presents results for the calculations with baseline epidemiological data using country-specific DWs. Denmark is not shown here, as no country-specific DWs were available. Sweden showed the largest change: the number of YLDs was 22% lower, resulting in a decrease in DALYs of 7%. For the other countries, the differences in YLDs and DALYs were smaller. Some changes appeared in the ranking of countries on DALYs, but only where the original differences in DALYs were small.

Figure 7.4. **Total DALYs, YLDs and YLLs for breast cancer by country (standardized)**



The burden of breast cancer IV: cross-national differences in epidemiological data and demographic data

DALYs computed from country-specific population figures, European DWs and the baseline epidemiological data by age are shown in table 7.2 (bottom). Total population size for each country was scaled to 100,000 to allow comparison. The number of DALYs increased for all countries: the difference ranging between 4.6 and 14.9% for Spain and Sweden, respectively. As a result, Sweden came fourth and France fifth in the ranking of the burden of breast cancer. The age pattern also changed and varied between countries, the number of DALYs still increased sharply towards age 50, but decreased more slowly thereafter (data not shown).

Discussion

The burden of breast cancer was compared in six European countries: Denmark, England and Wales, France, the Netherlands, Sweden, and Spain, and the relative effects of variation in epidemiology, in disability weights (DWs), and in demography on the estimated burden studied. Taking into account only

Table 7.1 **Breast cancer DALYs, YLLs and YLDs calculated for six European countries. Baseline calculation. Cross-national differences in epidemiological data.**

Baseline epidemiological data, European DW, standard EU population						
	Denmark	England and Wales	France	Netherlands	Sweden	Spain
DALY	1102.2	961.2	801.4	1071.0	745.1	741.9
YLL	868.8	768.5	606.4	818.2	520.9	526.7
YLD	233.4	192.8	195.0	252.8	224.1	215.3
%YLL	78.8	79.9	75.7	76.4	69.9	71.0

DWs: disability weights; EU: European; % YLD: percentage of DALYs that is attributable to morbidity.

Table 7.2 **Breast cancer DALYs, YLLs and YLDs calculated for six European countries. Additional calculations. Alternative epidemiological data, cross-national differences in DWs, cross national differences in demographic data**

	Denmark	England and Wales	France	Netherlands	Sweden	Spain
Alternative epidemiological data, European DW, standard EU population						
		% MI		*Mortality*		*Incidence*
DALY		961.7		1138.7		678.7
YLL		768.5		886.9		526.7
YLD		193.2		251.9		152.1
Change in comparsion with baseline calculation (%)						
DALY		0.0		6.3		-8.5
YLL		0.0		8.4		0.0
YLD		0.2		-0.4		-29.4
Baseline epidemiological data, country specific DW, standard EU population						
DALY		975.0	763.7	1120.2	695.1	764.1
YLL		768.5	606.4	818.2	520.9	526.7
YLD		206.6	157.3	302.0	174.2	237.4
Change in comparsion with baseline calculation (%)						
DALY		1.4	-4.7	4.6	-6.7	3.0
YLL		7.2	-19.4	19.5	-22.3	10.3
YLD						
Baseline epidemiological data, European DW, country-specific population figures (scaled)						
DALY	1219.0	1044.4	846.6	1125.3	856.2	776.2
YLL	958.2	830.2	637.3	857.2	595.9	556.7
YLD	260.8	214.2	209.3	268.1	260.2	219.5
Change in comparsion with baseline calculation (%)						
DALY	10.6	8.7	5.6	5.1	14.9	4.6
YLL	10.3	8.0	5.1	4.8	14.4	5.7
YLD	11.8	11.1	7.3	6.1	16.1	2.0

DWs: disability weight; EU: European.

differences in breast cancer epidemiology, relatively more DALYs were lost in Denmark and the Netherlands, followed by England and Wales. Spain lost the smallest number of DALYs to breast cancer. Cross-national variation in disease epidemiology was the largest source of variation in the burden of breast cancer. Additional variation resulting from introducing differences in population structure was smaller. The smallest effects resulted from uncertainty in the epidemiological data and using country-specific DWs.

The major part of the burden (at least 70%) was caused by premature mortality (YLLs). Consequently, the estimated burden of breast cancer reflected mainly mortality. To determine in which country the burden of breast cancer was largest, a simple inspection of the epidemiological frequency data would have sufficed: the differences in mortality rates could predict the rank order of the countries accurately. The use of a composite health indicator such as the DALY seems unnecessary to compare the problem of breast cancer between these European countries. Although this may seem surprising, it is in line with the rationale behind the construction of the DALY, which was to compare the burden of different diseases within one region, not of one disease between different regions.

DWs have raised a lot of debate regarding their reliability and comparability between countries [3]. The European disability weights project showed that variation in DWs between countries could only, in very small part, be explained by country-related effects and that the ranking of disease stages was very similar between countries [6]. The additional variation introduced using country-specific DWs (maximally 7%, Sweden) did not exceed the effect of uncertainties in the epidemiological data (6% for the Netherlands and 9% for Spain), and was smaller than the additional variation attributable to demography (up to 15%) and the variation resulting from cross-national differences in disease epidemiology (up to 33%). Therefore, using country-specific weights adds little additional information to comparison of the burden of breast cancer between these European countries. A similar conclusion was drawn with respect to the burden of dementia [6].

Even though we attempted to obtain reliable epidemiological data, these data are not perfect. Moreover, due to differences in registration practices, data source (expert opinion, hospital or cancer registration), coverage (regional versus national), case-definition and sometimes even the year of reference, comparability between countries may be limited. Nevertheless, the variation due

to uncertainty in the epidemiological data was small (maximally 9%, Spain). Breast cancer mortality data are believed to be rather reliable. In each country cause-specific mortality was registered on a regular basis with national coverage, increasing the comparability between countries. On the other hand, mortality data have a large influence on the burden of breast cancer; a small change often changed the ranking order between countries.

Comparability of the incidence data between countries is more limited, as we relied on different data sources (cancer registrations and hospital statistics) with different coverage (regional and national) and case-definitions (in situ tumours were included for England and Spain, but not for the other countries). Also, we were uncertain about the incidence estimates for Spain, because of the large discrepancy, especially in the age pattern, between the rates based on 1996 hospital discharge data and the older regional registration data. The former rates were higher at younger ages, which may be explained by a cohort effect of a trend of increasing breast cancer incidence, but can also be caused by overestimation resulting from referral shortly after diagnosis. Furthermore, for women over 65 the 1996 rates were lower than the older rates, which is unlikely considering the trend in incidence and the usually observed increase with age.

The data on the proportion of tumours metastasised at diagnosis (M1), of women undergoing mastectomy, and of secondary primaries (SP) were not readily available and therefore obtained from many different sources including expert opinion. Nevertheless, we believe these data have little effect on the estimated DALY. A different proportion M1 for England hardly affected the estimated burden. These data affect only the number of YLDs and only by altering the average DW through the distribution over the different disease phases. Likewise, the proportion of women undergoing mastectomy will have little effect. The percentage SP directly affects incidence, but, as it was small, its effect will also be small.

Conclusions

A large fraction of the burden of breast cancer was attributable to premature mortality (70% or more), causing this burden to reflect mainly mortality. To compare the burden of breast cancer in different countries a simple inspection of the mortality rates would have sufficed. As most types of cancers have a large mortality component we believe this conclusion can be extrapolated to most

other types of cancer. Using country-specific DWs will add little information to the comparison of their burden. However, for a disease that causes mainly morbidity, such as several psychiatric disorders, the effect of variation in DWs may be larger.

Acknowledgements

We thank Theo Vos (Public Health Division, Dept. of Human Services, Melbourne, Australia) for help and advice on epidemiology, modelling and burden of disease calculations and Joshua A. Salomon (WHO, Geneva, Switzerland) for general advice.
This study was supported by a grant from the BIOMED II programme of the European Union (project number BMH 4-98-3253).

Contributors

The burden of disease estimations for breast cancer were ultimately designed in the discussions of the European Disability Weights group. P.J. van der Maas, L.J. Gunning-Schepers, J Pereira, I. Durand-Zaleski, J. Raftery, F. Diderichsen and F. Kamper-Jørgensen were members of the Steering Committee of the European Disability Weights project. P.J. van der Maas acted as the Project Co-ordinator. The respective country teams collected the epidemiological and demographic data for each participating country. The writing Committee, consisting of M.E. Kruijshaar and J.J. Barendregt drafted the manuscript for this paper. All members of the European Disability Weights group contributed to the final results and to the final version of the manuscript.

Members of the European Disability Weights group:
Finn Kamper-Jørgensen, Ulla Christensen, Kim Moesgaard Iburg, (National Institute of Public Health, Copenhagen, Denmark); James Raftery, Claire Packer, Lisa Gold, Suzanne Robinson (from Oct 1999) (University of Birmingham, England); Isabelle Durand-Zaleski, Michael Schwarzinger (Public Health, Henri Mondor Hospital, AP-HP, Paris, France); Louise Gunning-Schepers, Gouke Bonsel, Clara Moerman (from Jan 2000), Marlies Stouthard (Academic Medical Center, University of Amsterdam, The Netherlands); Paul van der Maas (project co-ordinator), Marie-Louise Essink-Bot (Dept. of Public

Health, Erasmus University Rotterdam, The Netherlands); Joaquin Pereira, Ana Baylin (until Jan 1999), Eduardo Fernandez Zincke (National School of Public Health, Madrid, Spain); Finn Diderichsen, Kristina Burström, Rickard Ljung (Karolinska Institute, Stockholm, Sweden).

References

1. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge: Harvard University Press; 1996.
2. Ubel PA, Richardson J, Menzel P. Societal value, the person trade-off, and the dilemma of whose values to measure for cost-effectiveness analysis. *Health Econ* 2000;9(2):127-36.
3. Ustun TB, Rehm J, Chatterji S, Saxena S, Trotter R, Room R, et al. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* 1999;354(9173):111-5.
4. Jelsma J, Chivaura VG, Mhundwa K, De Weerdt W, de Cock P. The global burden of disease disability weights. *Lancet* 2000;355(9220):2079-80.
5. De Wit GA, Busschbach JJ, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? *Health Econ* 2000;9(2):109-26.
6. Essink-Bot ML, Pereira J, Packer C, Schwarzinger M, Burstrom K. Cross-national comparability of burden of disease estimates: the European Disability Weights Project. *Bull World Health Organ* 2002;80(8):644-52.
7. Stouthard MEA, Essink-Bot ML, Bonsel GJ, Barendregt JJ, Kramers PGN, van de Water HPA, et al. *Disability weights for diseases in the Netherlands.* Rotterdam, the Netherlands: Department of Public Health, Erasmus University Rotterdam; 1997.
8. Mathers CD, Vos ET, Stevenson CE, Begg SJ. The Australian Burden of Disease Study: measuring the loss of health from diseases, injuries and risk factors. *Med J Aust* 2000;172(12):592-6.
9. Melse JM, Essink-Bot ML, Kramers PG, Hoeymans N. A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *Am J Public Health* 2000;90(8):1241-7.
10. Peterson S, Backlund I, Diderichsen F. *Sjukdomsbördan i Sverige - en svensk DALY-kalkyl. [Burden of disease in Sweden - a Swedish DALY calculation].* Stockholm: National Public Health Institute (Folkhälsoinstitutet); 1998.

11. Mogren I, Stenlund H, Hogberg U. Long-term impact of reproductive factors on the risk of cervical, endometrial, ovarian and breast cancer. *Acta Oncol* 2001;40(7):849-54.
12. Murray CJ. Quantifying the burden of disease: the technical basis for disability- adjusted life years. *Bull World Health Organ* 1994;72(3):429-45.
13. WHO. *World Health Statistics Annual 1990*. Geneva: World Health Organization; 1991.
14. Hölzel D, Klamers A, Schmidt M. *Krebs: Häufigkeiten, Befunde und Behandlungsergebnisse; Persepektiven für die Krebsregisterdiskussion und eine quantitative klinisch-epidemiologische Onkologie aus dem Tumorregister München. [Cancer: data from the Munich cancer registry]*. München: W. Zuckschwerdt Verlag GmbH; 1996.
15. *XIXtrFun, cubic-spline*. J. Davita, netrax: www.netrax.net/~jdavita/XlXtrFun/XlXtrFun.htm, accessed 1999.
16. *DisMod II*. World Health Organization: www.who.int/m/topics/dismod/en/index.html, accessed 2001.
17. Barendregt JJ, Oortmarssen GJ van, Vos T, Murray CJL. A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. *Popul Health Metr* 2003;1(4):1-8.
18. Barendregt JJ, van Oortmarssen GJ, van Hout BA, van den Bosch JM, Bonneux L. Coping with multiple morbidity in a life table. *Mathematical Population Studies* 1998;7(1):29-49.
19. *EUCAN, database France, 1995*. IARC: www.iarc.fr/encr.html, accessed 2000.
20. *Clinical hospital discharges data set, Spain, 1997*. Spanish Ministry of Health and Consumers affairs: www.msc.es/cmbd/home.htm, accessed 2000.

ADDENDUM I

Description of phases of breast cancer.

The phases of breast cancer that were weighted in the European Disability Weights [1] project were provided with a lay text and a standardised generic description of the functional health status. For the generic description the Euroqol-5D (EQ-5D) classification system of health status (mobility, self-care, usual activities, pain / discomfort and anxiety / depression) was used and extended with a sixth dimension of cognitive functioning (EQ-5D+C) [6]. Each dimension had three levels of the general form 1 = no problems, 2 = some problems, and 3 = severe problems.

phase 1.

- Patient is initially diagnosed with breast cancer and undergoes breast-conserving surgery with subsequent local radiotherapy and chemotherapy, during a 6-month period. In the following 6 months the patient experiences some pain and discomfort from the wound. The surgery is intentionally curative, but there is the shock of the cancer diagnosis and uncertainty about lasting cure.
- EQ-5D+C during the 1st 6 months: no problems in walking about, no problems in washing or bathing, some problems in performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairment.

  EQ-5D+C during the 2nd 6 months: no problems in walking about, no problems in washing or bathing, no problems with performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairment.

phase 2.

- Patient is initially diagnosed with breast cancer and undergoes radical mastectomy (total removal of the breast and adjacent tissues), with or without subsequent radiotherapy and/or chemotherapy, during a 6-month period. In the following 6 months the patient experiences discomfort and pain from the wound (initially), disfigurement and scarring, and a swollen

arm. The surgery is intentionally curative, but there is the shock of the cancer diagnosis and uncertainty about lasting cure.

- EQ-5D+C during the 1st 6 months: no problems in walking about, some problems in washing or bathing, some problems in performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairment.

  EQ-5D+C during the 2nd 6 months: no problems in walking about, no problems in washing or bathing, some problems with performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairment.

phase 3.

- Patient who has undergone breast conserving therapy and local radiotherapy and/or adjuvant chemotherapy for breast cancer more than one year ago, who now only experiences some discomfort; there are no signs of tumour recurrence.
- EQ-5D+C: no problems in walking about, no problems in washing or bathing, no problems in performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairments.

phase 4.

- Patient who has undergone a radical mastectomy with or without radiation therapy and/or chemotherapy for breast cancer, more than one year ago. The permanent after effects consist of discomfort and pain from disfigurement and scarring, and a swollen arm. There are no signs of tumour recurrence.
- EQ-5D+C: no problems in walking about, no problems in washing or bathing, some problems in performing usual activities, some pain or discomfort, moderately anxious or depressed, no cognitive impairments.

phase 5.

- Patient with breast cancer, who has local recurrence of the disease and/or spread elsewhere (to the bone, brain or liver), requiring chemotherapy and/or radiotherapy. The patient spends a large amount of time on treatment and experiences anxiety and major discomfort (e.g., due to side effects of chemotherapy).

- EQ-5D+C: some problems in walking about, some problems in washing or bathing, unable to perform usual activities, some pain or discomfort, very anxious or depressed, no cognitive impairments

Weights derived in the European Disability Weights project

In the European Disability Weights project, Disability Weights (DWs) were derived from health and non-health professionals using time trade-off (TTO), person trade-off (PTO) and a visual analogue scale (VAS). Because derivation of VAS scores does not include an explicit trade-off, VAS scores are not considered suitable for direct use as disability weights. Disease stages for which only VAS scores were available were transformed to TTO scores using: $TTO = (1-(1-VAS)^{\wedge}\alpha)$ [6], in which α differs between countries. The resulting weights are given in Table 7A.I.

Table 7A.I. **European and country-specific disability weights.**

Phase of breast cancer	country					
	EU	E&W	Fr	NL	Sp	Sw
- diagnosis and primary breast conserving therapy[a]	0.27	0.32	0.21	0.30	0.32	0.19
- diagnosis and radical mastectomy[a]	0.37	0.38	0.37	0.41	0.47	0.26
- clinically disease-free without major permanent sequelae[b]	0.18	0.19	0.15	0.20	0.23	0.12
- clinically disease-free with major permanent sequelae[a]	0.30	0.32	0.22	0.35	0.35	0.22
- recurrent or disseminated disease[a]	0.61	0.60	0.64	0.70	0.58	0.53

EU: the pooled European weights, obtained by pooling the data from the five countries
[a]: VAS scores were obtained and transformed using a transformation formula
[b]: TTO scores were obtained

ADDENDUM II

Table 7A.II **A. Epidemiological frequency data: data source and reference year by country.**

Country	Data	Year	Source
Mortality	*by 5-year age-groups up to 85+*		
Denmark	Numbers of cause-specific deaths	1996	Danish Cause of Death Registration, National Board of Health; (NBH) Denmark
England&Wales	Numbers of cause-specific deaths	1996	Office for National Statistics (ONS)
France	Numbers of cause-specific deaths	1996	National Institute of Health and Medical Research France;
	Numbers of prior year (alternative)	1995	National Institute of Health and Medical Research France
Netherlands	Numbers of cause-specific deaths	1996	Statistics Netherlands (CBS)
Spain	Numbers of cause-specific deaths	1996	National Institute of Statistics Spain (INE)
Sweden	Numbers of cause-specific deaths	1996	National Board of Health and Welfare Sweden (Socialstyrelsen)

Table 7A.II **A. Continued**

Incidence	*by 5-year age-groups up to 85+*		
Denmark	Numbers of invasive tumours	1996	Danish Cancer Registry, NBH Denmark;
	% SP	1995-1997	No data available, NL estimate used after expert consultation
England&Wales	Rates, including non-invasive	1994	Office for National Statistics (ONS);
	% SP	1995-1997	No data available, NL estimate used after expert consultation
France	Numbers of invasive tumours	1995	EUCAN[1] ;
	% SP	1995-1997	No data available, NL estimate used after expert consultation
Netherlands	Numbers of invasive tumours	1996	the Netherlands Cancer Registry (NKR);
	% SP	1995-1997	Comprehensive Cancer Centre Amsterdam (IKA)
Spain	Numbers of new discharges	1996	Hospital morbidity survey (from INE);
	Rate invasive tumours (alternative)	1986-92[2]	Nine regional registrations[2];
	% SP	n.a.	No data available, experts estimated SP as close to zero
Sweden	Numbers of invasive tumours	1996	National Board of Health and Welfare Sweden (Socialstyrelsen)
	% SP	1995-1997	No data available, NL estimate used after expert consultation

alternative: used as additional data source, because different years or sources showed very different patterns; % SP: percentage of secondary breast cancer primaries; n.a.: year is not applicable, as the estimates concern expert opinions, NL: estimates from the Netherlands were used.
[1] incidence numbers (to 75+) were estimated from regional incidence data (around 1990), mortality (1994) and population figures (1995) [19].
[2]: Zaragoza 1986-1990, Tarragona 1988-1992, Navarra 1987-1991, Murcia 1988-1992, Mallorca 1988-1992, Granada 1988-1992, Basque Country 1988-1991, Asturias 1988-1991, Albacete 1991-1992. Average age-specific rates were weighted for total population size of each registration area.

Table 7A.II **B. Epidemiological frequency data: data source and reference year by country.**

Country	**Data**	**Year**	**Source**
Demography	*by one-year age-groups*		
Denmark	Population numbers,	1996	Statistics Denmark;
	total mortality	1996	Danish Cause of Death Registration, NBH Denmark
England&Wales	Population numbers,	1996	Office for National Statistics (ONS);
	total mortality	1996	Office for National Statistics (ONS)
France	Population numbers,	1996,1995	National Institute of Statistics and Economic Studies France;
	total mortality	1996	National Institute of Statistics and Economic Studies France
Netherlands	Population numbers,	1996	Statistics Netherlands (CBS);
	total mortality	1996	Statistics Netherlands (CBS)
Spain	Population numbers,	1996	National Institute of Statistics Spain (INE);
	total mortality	1996	National Institute of Statistics Spain (INE)
Sweden	Population numbers,	1996	Statistics Sweden (SCB);
	total mortality	1996	National Board of Health and Welfare Sweden (Socialstyrelsen)

Table 7A.II **C. Additional epidemiological data: data source and reference year by country.**

Country	Data	Year	Source
Additional data	*all age-groups combined*		
Denmark	%M1	1990-1995	Danish Cancer Registry, NBH Denmark;
	%mastectomy	1990-1995	Danish Cancer Registry, NBH Denmark
England&Wales	%M1	1996	West Midlands Regional Cancer Registry;
	% M1 (alternative)	n.a.	Expert opinion;
	%mastectomy	1998-1999	Hospital episode statistics, ONS
France	%M1	n.a.	Expert opinion;
	%mastectomy, ages <70 and >70	<=1998	Database of hospital admissions, INSEE
Netherlands	%M1	1995-1997	Comprehensive Cancer Centre Amsterdam (IKA);
	%mastectomy	1995-1997	Comprehensive Cancer Centre Amsterdam (IKA)
Spain	%M1	1996	Granada Regional Cancer Registry;
	%mastectomy	1997	Clinical hospital discharges data set [20]
Sweden	%M1	n.a.	Expert opinion
	%mastectomy	n.a.	Expert opinion

%M1: proportion of incident cases that are metastasised at diagnosis; % mastectomy: proportion of women undergoing mastectomy; n.a.: year is not applicable, as the estimates concern expert opinions; alternative: used as additional data source, because cancer epidemiologists suspected the data on the %M1 to be low. alternative: used as additional data source, because different years or sources showed very different patterns.

ADDENDUM III

General approach

Different phases of breast cancer are weighted differently. In order to calculate Years Lived with Disability (YLDs) using the incidence method, data on the incidence and duration of each of these phases are needed. Therefore we used a model that describes the possible phases that patients may pass through over time. This model is depicted in Figure 7A.III. It distinguishes five different pathways: for patients who will survive their diagnosis after breast-conserving therapy (A), for patients who will survive their diagnosis after mastectomy (B), for those who will not survive their diagnosis after breast-conserving therapy, but do not have metastases at diagnosis (C), for those who will not survive their diagnosis after mastectomy, but do not have metastases at diagnosis (D), for those who have metastases at diagnosis and do not survive (E). For all women without metastasis at diagnosis the phase of diagnosis and therapy is the starting point and is followed by a disease-free phase. This phase is split into disease-free with and without sequelae using the percentage undergoing mastectomy treatment, because we assumed that breast conserving therapy results in fewer and less severe sequelae than mastectomy. Non-survivors progress to the phase of metastasis or local recurrence after the disease-free phase, while for survivors it is the end stage. As we assumed women return to a state of full health after five years of disease-free survival following breast-conserving therapy, and to a more optimal state (approximated by the difference between disease-free with and without sequelae) following mastectomy, we added phase 6 and 7 which represent the extension of the disease-free phase after five years. M1 women enter the phase of metastasis or local recurrence immediately at diagnosis.

Estimation of phase-specific incidence

In order to distribute incidence over these pathways, we distinguished survivors from non-survivors, tumours metastasised at diagnosis (M1) from those not metastasised at diagnosis, and cases undergoing mastectomy from breast-conserving treatment. We assumed that all non-survivors in the model die of breast cancer and that this was always preceded by metastasis or local recurrence. The proportion of survivors ("the proportion cured") was estimated

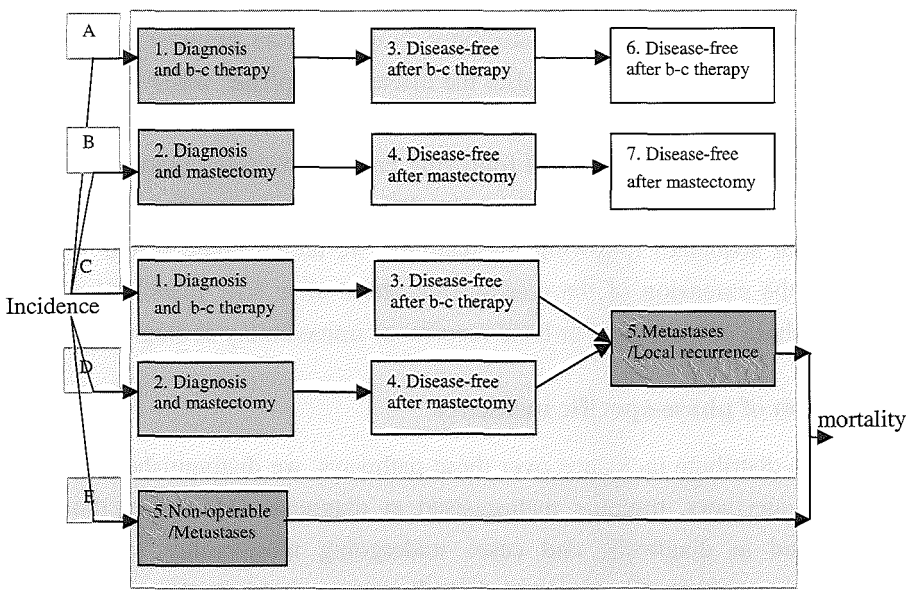
from the ratio between mortality and incidence. Because death resulting from incidence at age n may occur many years later, mortality at age n can not be contributed to incidence at age n. Therefore, the "proportion cured" can not be calculated age-specifically using this ratio. We estimated the proportion cured using:

$c = 1 - \left[\sum_n (mortality_n * EUpop_n) \; / \sum_n (incidence_n * EUpop_n)\right]$, where c is the proportion of breast cancer survivors, *mortality* and *incidence* are the mortality and incidence rates at age group n and *EUpop* the European standard population at age group n. Incidence was then split over the five pathways using the data on the proportion of women undergoing mastectomy, *prop_mastec*, and the proportion M1, *prop_M1*:

$incidence_{n,A} = incidence_n * c * (1 - prop\_mastectomy)$,

$incidence_{n,B} = incidence_n * c * prop\_mastectomy$,

Figure 7A.III. **Different pathways that patients may follow after diagnosis.**



B-c therapy: breast-conserving therapy.

$$incidence_{n,C} = [incidence_n * (1 - c - prop\_M1)] * (1 - prop\_mastectomy),$$

$$incidence_{n,D} = [incidence_n * (1 - c - prop\_M1)] * prop\_mastectomy,$$

$$incidence_{n,E} = incidence_n * prop\_M1.$$

Estimation of phase-specific duration

The estimates of duration and how they were calculated are shown per phase in Table 7A.III. For non-survivors duration was assumed independent of general mortality, because of the assumption that all non-survivors die from breast cancer. Survivors, nevertheless, die from other causes of mortality, and for them duration depends on general mortality. The duration of phase 6 and 7, are therefore calculated from the average total disease duration calculated from DisMod II [16], which takes death from other causes into account. At older ages the duration of phase 3 and 4 may decrease because general mortality increases.

Table 7A.III. **Duration estimates and source of the estimates.**

Phase {pathway}	Duration *d*		Source
1, 2 {A, B, C, D}	1 year		Imposed by description.
3, 4	5 year;	for *d* 6,7 > 0	A matter of choice;
{A, B}	5 year – *d* 6,7;	for *d* 6,7 < 0	Subtract general mortality
3, 4 {C, D}	24 months		Munich cancer registry [14]
6, 7 {A, B}	[*dtotal* – (1– *c*– *prop_M1*) * (*d* 1,2 + *d* 3,4 {C,D} + *d* 5) – *prop_M1* *d* 5] / c – d 1,2 – d 3,4 {A,B}		*dtotal* from DisMod
5 {E}	34months – d 1,2 = 22 months		Munich cancer registry [14]

d 1,2; *d* 3,4; *d* 5 is duration of phase 1 and 2 ; 3 and 4; and phase 5, respectively.
dtotal is the average total disease duration estimated from DisMod II.
c is the proportion cured.

8

General discussion

Introduction

In Chapter one we introduced the research questions of this thesis. We argued that, as health policy decisions are mostly taken at the level of diseases and risk factors, disease-specific summary measures of population health are more useful than generic ones, at least in theory. In practice their use may be limited by their extensive data requirements. To calculate these measures two types of data are required, both of which pose specific problems to the use of disease-specific summary measures. On the one hand disease-specific epidemiological frequency data are needed. Unfortunately, these data are not always readily available, and can be of dubious quality. On the other hand disease-specific health status valuations ("disability weights" in the DALY methodology) are required to weigh the years lived with a disease. These weights should refer to the same severity level of the disease as the epidemiology data, which is not always straightforward to achieve.

Two specific research questions were formulated:

1) To what extent are causal disease models valid and useful to check the consistency of epidemiological frequency data and to supplement them?
2) How can disease-stage specific health status valuations be tailored to epidemiological frequency data?

With respect to these two research questions, the main findings of this thesis were:

1) Causal disease models can be used to detect consistency problems in the epidemiological data and to supplement them, but time trends complicate their use.
2) Different methods can be used to tailor stage-specific health status valuations to the epidemiology.

Overall, we concluded from this that it is possible to construct useful disease-specific summary measures of population health in practice. These findings will be discussed in the following sections, and focussing mainly on the DALY measure. We start with reporting the main findings of this thesis. Next we address the two research questions. Finally, we will draw some conclusions and give some recommendations.

Main findings

In part A of this thesis we studied the validity and usefulness of disease models. The main findings of this part were:

1. The validity of incidence-prevalence-mortality (IPM) models was supported by a check against artificial data.
2. Both data inconsistencies and trends can cause model outcome and data to be discrepant. For breast cancer, a large part of the discrepancy between model outcome and data could be explained by underestimation of the prevalence data and a trend of increasing incidence.
3. Application of a disease model to data for Major Depression (MD) required a complex model, but allowed the estimation of lifetime prevalence, a parameter that is difficult to measure.

Part B concerned the tailoring of disability weights (DWs) to epidemiological data. The main findings of this part were:

1. On the basis of disability data three clusters of MD could be distinguished: "mild" MD, "moderate to severe" MD and MD "with psychotic features".
2. Newly derived DWs for these three clusters, and the overall DW in which they are combined, were very similar to other studies. One exception was the overall DW used in the Dutch Burden of Disease Study, which was 73% lower due to a different estimate of the proportional prevalence of the clusters.
3. The estimated burden of breast cancer was most sensitive to cross-national variation in disease epidemiology. Variation in disability weights and uncertainty in epidemiological data had smaller effects.

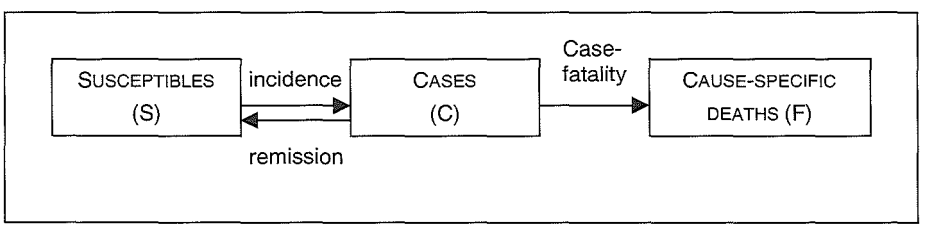
Usefulness of disease modelling

To get around some of the problems with the availability and validity of the epidemiological data, disease models have been developed that formalise the relations between the different epidemiological parameters. These models can be seen as a form of bookkeeping of individuals, as they keep track of all the transitions that individuals make between states in the natural course of a disease process. Figure 8.1 shows that in the natural course of a disease, incidence always has to precede prevalence, and mortality can only follow from having the disease. Thus, when an initially disease-free cohort is followed over time and all transitions from healthy to diseased (incidence), diseased to death (case-fatality)

and diseased to healthy (remission) are taken into account, the number of cases (prevalence) can be calculated. This simple way of bookkeeping forms the basis of IPM models (incidence, prevalence, mortality models). These models, of which the DisMod model [1, 2] is an example, have been used both to calculate missing data and to check the internal consistency of data [3-5]. When for example prevalence data are not available, it can be calculated from incidence, remission and cause-specific mortality data. When prevalence is also available and applied to the model, inconsistencies between the data will appear as discrepancies between model outcome and data. In this way inconsistencies can be detected and one can choose to adjust the observed data for the inconsistencies. IPM models use only three parameters, with remission sometimes being zero. The validity and usefulness of these models in checking the consistency of data and supplementing them is subject of the first part of this thesis.

We studied the validity and usefulness of IPM models by testing them against data sets for different types of cancer: two complete and consistent artificial data sets and four high quality empirical data sets. The artificial data were used to check the appropriateness of the Markov assumption in the models. Like every model, IPM models are a simplification of reality. One of the assumptions that is made in simple IPM models like DisMod is the Markov assumption of a random process whose future probabilities are determined by its most recent values [6, 7]. In other words, in the disease model the probability of transition to a next state is assumed to be determined only by the value of the previous state. The advantage of this premise is that the model does not have to keep track of transitions in the past: there is no need for memory in the model.

Figure 8.1. **The natural course of a disease**



However, this also means that the often-occurring relationship between mortality from a disease and the time lived with this disease can not be taken into account. We therefore tested the model against artificial data sets that were generated by a model that does include time-dependence of cancer mortality. This model was constructed and extensively tested for the evaluation of cancer screening programmes (MISCAN) [8]. Applying the artificial data to the IPM models gave no evidence that the Markov assumption was invalid at a population level: the IPM models reproduced the artificial data very well.

To test the usefulness of IPM models in practice, we next used four presumably high quality empirical data sets. The results of this analysis casted some doubts on their usefulness: even though empirical data for different types of cancer are of high quality, large discrepancies between measurements and calculations occurred. It is well known that not only data inaccuracies, but also past trends in incidence or mortality can cause such discrepancies. Time trends may cause the data to appear inconsistent in steady state models, while in fact they are not. This is caused by the fact that prevalence is a stock variable: it is the accumulation of incident cases from the past. As a result, it cannot react instantaneously to changes in incidence and case-fatality, but only with a certain delay. Although it is possible to account for the effects of time trends in a dynamic model, this requires additional input data on the nature and size of the trends, which are not available for most diseases. Often, we do not even know whether a trend is present or not, and the researcher faces a dilemma what to do with the discrepancies. Adjusting observed data for apparent inconsistencies that are in fact the consequence of past trends would rather defeat the purpose of IPM models.

We tried to throw some light upon this dilemma using the fact that for breast cancer in the Netherlands there is a tentative estimate of the trend in incidence [9] and several of the data problems are known. With this information we quantified the relative contribution of trends and known data problems on the discrepancy for breast cancer. Two factors had a major effect in this analysis: an underestimation of the 1993 prevalence data from the Regional Cancer Centre South (IKZ), and a trend of increasing incidence. Together, they accounted for a major part (60%) of the discrepancy. The underestimation of the prevalence results from the fact that before 1970 incident cases were not registered, and could thus not be included in the prevalence figures. Although prevalence of breast cancer is not a very useful epidemiological measure and this particular

underestimation will furthermore decrease when time progresses, this analysis is important as it shows the ability of the IPM models to detect data problems. While it is a known fact that the registration in the Southern part of the Netherlands was complete since 1970, the effect of this on the prevalence data for breast cancer had not been assessed earlier.

The results of the analyses, however, also indicated that trends can have a large effect on the model results. The trend in incidence could have such a large effect on the model for breast cancer by the combination of two factors. First, the trend was present for a long time allowing the increase in incidence to build up. Second, the outflow from prevalence to mortality is not very fast for breast cancer, causing prevalence to reflect a range of incidence rates from the present to many years ago. The effect of trends for diseases with a high case-fatality rate will thus be smaller. Nevertheless, even for these cases the problem remains that this effect can not be quantified unless there are data on the size and nature of the trend. Because, unlike for breast cancer, for most diseases these data are not available, there always remains the need for careful interpretation of the modelling results.

After applying the models to these high quality data for breast cancer, we confronted them with the less well monitored epidemiology of Major Depression (MD). Prevalence and incidence data for MD in the Netherlands were available from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). In this study the information was obtained with reference to several time frames. Prevalence was assessed in the past month, year and life (lifetime prevalence), first-incidence (of a first episode) in the past life, and total incidence (both first and recurrent episodes) in the past year. Simple IPM models, nonetheless, can solely link epidemiological parameters referring to one time frame at a time. Sufficient data were available only to model the lifetime perspective, as the study provided no data on remission (remission from lifetime prevalence is by definition zero). Lifetime prevalence measured in a cross-sectional way may, however, be largely underestimated by recall problems [10, 11]: to obtain lifetime prevalence in a cross-sectional surveys, respondents have to recall the presence and timing of symptoms retrospectively over their past life. A simple inspection of the NEMESIS lifetime data suggested that recall bias could indeed be present in these data. Examining the data by age showed a decrease after age 45 that could be due to recall bias, although also excess

mortality and trends could explain a decrease at older ages. A quick application of the lifetime data to DisMod furthermore showed that they were only consistent with an unrealistically high excess mortality that far exceeded the relative risks (RR) reported in the literature [12]. Despite the fact that, once more, DisMod identified a data problem, IPM models can not help to improve the epidemiological estimates of MD: the lifetime data were unreliable, and there were insufficient data to model other time frames.

For this reason we applied a more complex disease model to Major Depression. We chose to construct a model using a microsimulation approach. This is a flexible technique that allows the use of parameters from different time frames as input and is able to take into account the large heterogeneity of MD. Microsimulation describes the disease process of an individual in terms of probabilities and their distributions [13]. Individual life histories are generated by random drawings from these distributions. Adding a large number of life histories creates a population from which community-based epidemiological measures, such as lifetime prevalence and number of episodes, can be derived. Using a microsimulation model we estimated lifetime prevalence of MD from the NEMESIS month and year prevalence data. The results pointed at two causes that result in an underestimation of the NEMESIS lifetime prevalence data: a cross-sectional setting and recall problems. To start with the first, the model estimates that approximately 30% of men and 40% of women suffer from one or more episodes of MD during their life. In a cross-section of age-groups similar to NEMESIS, nevertheless, only 20% and 30% of the modelled life histories were MD positive. The potential for recall bias is then indicated by the difference between these cross-sectional estimates and the NEMESIS data, which were 38% lower. The recall bias in these data may thus be considerable, posing questions at the usefulness of the empirical measurement of lifetime prevalence. Again, models appear to be useful, this time not only in consistency checking, but also in supplementing data: the model also provided us with an (indirect) estimate for a study in which this estimate was not measured empirically (the Australian Mental Health and Wellbeing Survey). This modelling exercise furthermore showed that for the complex epidemiology of depression a more complex disease model was required. It is likely that simple IPM models will be inadequate also for other diseases with a complex epidemiology, such as other psychiatric disorders, with problems similar to MD,

or several diseases of the locomotor system (e.g. low back pain and RSI). However, following the principle of parsimony (Occam's razor) these models should only be used when simple IPM models are not sufficient.

Disability, disability weights and burden

Disease-specific health status valuations, or Disability Weights (DW) in the DALY methodology, should be tailored to the epidemiological case-definition: to prevent over- and underestimation of the summary measure both should refer to the same severity level of a disease. The tailoring of DWs is facilitated by some refinements that were made to the original DW methodology of the Global Burden of Disease Study (GBD) by a Dutch group of researchers. In the Dutch Disability Weights Study (DDW) [14] researchers added a formal health status description in EuroQol to each health state and, where previously only one weight was available per disease, they derived weights for several distinguished disease-stages. The formal health status description gives an explicit indication of the severity level of the disease (or disease-stage), which before was implicit. This information can be used to match DWs and epidemiology (if health status information is also available for the epidemiological data). For example, in the Australian Burden of Disease Study disability information (SF-12) that was collected together with the prevalence data for MD was used to match DWs and epidemiology [4].

The stage-specific weights can be combined into a range of disease-specific weights, by applying different estimates of the proportional prevalence of the stages. As a result, a disease-specific DW can be adapted in response to improvements in therapy or to a change in the used diagnostic instrument, or case-definition. However, the division into stage-specific weights also has a downside. To combine them into an overall DW for a specific disease, information is needed on the proportional distribution of the disease across the stages. This information is difficult to obtain. Epidemiological data are often not available per disease-stage, and if they are, the disease-stage descriptions used in epidemiology may differ from the ones used in the valuation. For example, the TNM-staging used in cancer epidemiology is not adopted here to distinguish the health-states for valuation, because they do not provide sufficiently different levels of health. How to tailor the stage-specific DWs to the epidemiological data was our object of study in part two of this thesis. Also, some attention was

given to the sensitivity of the summary measure to DWs, to assess the importance of this tailoring to the overall summary measure.

For Major Depression (MD), data from the Netherlands Mental Health Survey and Incidence Study (NEMESIS) provided a good opportunity to tailor DWs to the epidemiology. The NEMESIS database contains information on both epidemiology and health status/ disability by different classes (type and severity) of MD. From the DDW study [14] weights for depression were available per severity class. However, the weights from this study seem to be oriented towards a clinical setting rather than a population, as the provided health status descriptions were based on expert opinion (medical doctors) and diagnostic criteria. Because the weights have to be combined with epidemiological data from population surveys, and these may include milder cases of depression than those found in clinical settings [15], the DDW weights do not match the epidemiology. Therefore, tailoring the DWs for MD requires both: 1) deriving new stage-specific DWs for tailored health status descriptions, and 2) combining the weights into an overall DW using a tailored proportional distribution of the disability levels.

Therefore we first set out to construct stages and their descriptions, and derive DWs. Using the NEMESIS data we examined how disability differed by severity class and type (single or recurrent) of MD. Only three levels of disability were distinguished: "mild", "moderate to severe" and "severe with psychotic features". This shows that type of MD is not needed to describe different disability levels, and that only three severity classes are needed, one less than in the DDW study. The disability information per disease-stage was less severe in NEMESIS than the descriptions used in the DDW, but the new stage-specific DWs for these descriptions were not significantly different from DDW weights. Apparently, the disease-stage label is much more important than the health status description in the health status valuation. This could be used as an argument against including the disease label in the health status description. Omitting this label, would ensure that what is valued is the health status description. However, one may also reason that the disease label is very important. It adds information to the formal description on domains that are not represented in the EuroQol system. Without a disease label the DW could be applied to another disease-stage with similar EuroQol levels, and would thus no longer be disease-specific. For this reason we believe the disease label should be included, although other people may argue differently.

In the second part of the tailoring we combined the new stage-specific weights into an overall DW using the proportional distribution from NEMESIS. Because this distribution was very different from the one used previously in the Netherlands (based on expert opinion, Dutch 1994 National Burden of Disease Calculation [16]), our resulting overall DW was very different (73% higher). This shows the importance of the quantitative epidemiological data in the burden of disease calculations. The overall DW for depression seems more sensitive to the distribution of disability across the depressed population, than to differences in the descriptions of the health status. Unfortunately, information on the proportional distribution is often lacking. This study shows one way to derive such information, but depends on the availability of a database with both epidemiological data and information on disability and stages.

For breast cancer, no such database was available. Epidemiological data for types of cancer are registered according to stages (TNM classification) that are very different from the ones for which DWs are derived (phases in the course of the disease). Thus, information on the proportional distribution has to be derived by other means than using the epidemiological databases. We used a modelling approach developed by Melse et al. [17, 18] that was also used in the Australian Burden of Disease Study [4, 19]. The model describes different pathways that incident breast cancer patients may pass through time. Each pathway consists of one or more of the discerned phases. Using some additional information to divide incidence over the different pathways (e.g. proportion of different therapy given in each country) and information on the duration of the phases [20] it was possible to quantify the proportional distribution across the phases.

Using this information we calculated the burden of breast cancer in six different European countries and assessed its sensitivity to different sources of variation. The major part of the burden (more than 70%) was caused by premature mortality. As a result, the highest burden of breast cancer was found in those countries that reported the highest mortality rates, casting doubt on the usefulness of a composite health indicator such as the DALY to compare diseases with a high mortality component. Not surprisingly, the estimated burden was also most sensitive to cross-national variation in disease epidemiology (read: mortality), while the country-specific DWs had the smallest impact. We believe this conclusion can be extrapolated to many other diseases,

as many diseases have a large mortality component. For these diseases the influence of DWs will be small.

On the other hand, for a disease that causes mainly morbidity, such as depression, the effect of variation in DWs may be larger. However, even for depression the effect of the DWs on the overall summary measure will be limited, because of the value of the DW itself. For disorders that induce mainly morbidity, the relative effect that variation in the DWs has on a summary measure depends on the size of the DWs. In the severe range of disability the relative variation resulting from, for instance, a difference of 0.1 is much smaller than in the mild range of disability. For example, adding 0.1 to the DW for depression (DW = 0.46, see chapter five) infers a change of 20% while for visual impairments (DW = 0.1 [16]), the difference is 100%. Because some diseases with a low DW are highly prevalent, the effect of variation in this DW on a population measure may be extensive. The major concern of tailoring the DWs is thus for highly prevalent diseases with a relatively low DW.

Conclusions

The first part of this thesis addressed the question: to what extent are causal disease models valid and useful to check the consistency of epidemiological frequency data and to supplement them. IPM models were shown to be valid and useful: even in the case of the high-quality data for breast cancer data problems could be detected. However, inconsistencies may also be caused by trends in incidence and/ or mortality, requiring careful interpretation of the result of models. Furthermore for some disorders, like depression, more complex models are needed. These also proved useful, both to detect data inconsistencies and to supplement missing data. We therefore conclude that causal disease models can be used to detect consistency problems in the epidemiological data and to supplement them, but that the presence of trends complicates their use.

The question: how can disease-stage specific health status valuations be tailored to epidemiological frequency data, had more than one answer. If available, epidemiological databases that contain both information on epidemiology and health status/ disability can be used both to provide a tailored disease-stage description (to derive new disability weights if desired) and to estimate the proportional distribution of the stages. In the case of depression such data were available and used. For breast cancer such data were not

available, and instead the proportional distribution was estimated using a modelling approach. Thus, there are different ways to tailor stage-specific health status valuations to the epidemiology of a disease.

In conclusion, useful disease-specific summary measures can be constructed in practice. Modelling provides a useful and valid tool to tackle some of the problems concerning the epidemiological data. Also, several ways were found to tailor the disability weights (DWs) to the epidemiological data. Nevertheless, the disease-specific summary measures are not very precise and should not be interpreted as such. They are useful to policy makers because they provide a ranking of diseases taking into account both morbidity and mortality and, furthermore, identify gaps in knowledge. The estimates should, however, not be interpreted as more than an indication of the importance of the disease.

Recommendations

Our research stressed the importance of the epidemiological data in the calculation of these disease-specific summary measures. Research in this field should therefore focus on obtaining sound epidemiological data more than on the DWs. Improvement of the epidemiological data input in summary measures could best be effectuated using both modelling and data collection. Until now, models have not been adopted frequently. The application of models to specific areas can improve the epidemiological input, as shown in the example of the depression model. More and better data collection, however, remains the most important, as models can not answer all the questions, especially when trends are present. Registration of the available information in a central database that is regularly updated would greatly facilitate the construction of summary measures and also the assessment of health impacts. In part, such a database already exists in the Netherlands at the Netherlands Institute of Public Health and the Environment (department of Public Health Status and Forecasting), but it could be extended to include more studies, more diseases, and the results of modelling.

The DWs were shown to be often a less important source of uncertainty in the calculation of disease-specific summary measures. With respect to the impact of the DWs on the summary measure, three groups of diseases can be discerned. The first group, diseases with a large mortality burden (like breast cancer), is relatively insensitive to variation in DWs. The second category, diseases causing mainly morbidity and having a DW in the severe range of

disability (such as depression), is sensitive to variation in the proportional distribution, but the relative variation in the stage-specific DWs remains limited. For diseases of these two groups, the relative variation in the DWs thus has limited consequences (unless different methods to derive the weights are applied), and research should focus mainly on the proportional distribution of the disease-stages. The inclusion of disability measures, such as the EuroQol, in national epidemiological surveys (such as for example in the second Dutch National Study on General Practice) would greatly facilitate the estimation of the proportional distribution. On the other hand, the third group of diseases is both sensitive to variation in the distribution as well as to variation in the stage-specific DWs. This group encompasses high prevalent disorders that produce mainly mild morbidity. Having a low DW, the relative variation in these weights may be large, while on the low side of the disability range the derived DWs may be imprecise [14]. For this type of diseases attention should therefore also be given to improving the stage-specific DWs themselves. More reliable DWs may be obtained by deriving them in relation to another mild to moderate disease instead of in relation to optimal health.

Finally, the abundance of diseases will restrict the practical application of disease-specific summary measures. Tailoring DWs and modelling disease epidemiology is time-consuming work that can only be done for a limited number of diseases. The intended use of the measure should therefore determine how much effort should be put in obtaining consistent epidemiological data and DWs. For simple comparison goals, generic summary measures may suffice, while the evaluation of interventions will require disease-specific information for some specific diseases at least. In the same line, there may be good reasons to adapt the design of the summary measure to the aim of its use. Monitoring health states in developing countries, for example, using a gap measure based on the same ideal life expectancy as in Western countries, will obscure small improvements in life expectancy. It may even be proposed to return to use the infant mortality rate for this kind of purpose [21]. Summary measures can have many different designs and the researcher should carefully choose the one that best suits the intended use.

References

1. Barendregt JJ, Oortmarssen GJ, van, Vos T, Murray CJL. A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. *Population Health Metrics* 2003;1(1):4.
2. Murray CJ, Lopez AD. Quantifying disability: data, methods and results. *Bull World Health Organ* 1994;72(3):481-94.
3. Barendregt JJ, Baan CA, Bonneux L. An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus. *Epidemiology* 2000;11(3):274-279.
4. Mathers CD, Vos T, Stevenson C. *The burden of disease and injury in Australia.* Canberra: AIHW; 1999. Report No.: PHE 17.
5. Murray CJL, Lopez AD, editors. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Cambridge, MA: Harvard School of Public Health; 1996.
6. Bharucha-Reid AT. *Elements of the Theory of Markov Processes and Their Applications.* New York: McGraw-Hill; 1960.
7. Papoulis A. "Brownian Movement and Markoff Processes." In: *Probability, Random Variables, and Stochastic Processes.* 2nd ed. ed. New York: McGraw-Hill; 1984. p. 515-553.
8. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. *Comput Methods Programs Biomed* 1985;20(1):79-93.
9. Coebergh JW, Crommelin MA, Kluck HM, van Beek M, van der Horst F, Verhagen-Teulings MT. Breast cancer in southeast North Brabant and in North Limburg; trends in incidence and earlier diagnosis in an unscreened female population, 1975-1986 [in Dutch]. *Ned Tijdschr Geneeskd* 1990;134(15):760-5.
10. Andrews G, Anstey K, Brodaty H, Issakidis C, Luscombe G. Recall of depressive episode 25 years previously. *Psychol Med* 1999;29(4):787-91.
11. Angst J. Epidemiology of depression. *Psychopharmacology (Berl)* 1992;106 Suppl:S71-4.
12. Cuijpers P, Smit F. Excess mortality in depression: a meta-analysis of community studies. *J Affect Disord* 2002;72(3):227-36.
13. Habbema JD, De Vlas SJ, Plaisier AP, Van Oortmarssen GJ. The microsimulation approach to epidemiologic modeling of helminthic infections, with special reference to schistosomiasis. *Am J Trop Med Hyg* 1996;55(5 Suppl):165-9.
14. Stouthard MEA, Essink-Bot ML, Bonsel GJ, Barendregt JJ, Kramers PGN, van de Water HPA, et al. *Disability weights for diseases in the Netherlands.* Rotterdam, the Netherlands: Department of Public Health, Erasmus University Rotterdam; 1997.

15. Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, et al. Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy. *Arch Gen Psychiatry* 1998;55(2):109-15.
16. Melse JM, Essink-Bot ML, Kramers PG, Hoeymans N. A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *Am J Public Health* 2000;90(8):1241-7.
17. Melse JM, Kramers PGN. *Berekeningen van de ziektelast in Nederland. Achtergronddocument bij VTV-1997; deel III, hoofdstuk 7.* Bilthoven: Rijksinstituut voor Volksgezondheid en Milieu (RIVM); 1998. Report No.: 431501028.
18. van der Maas PJ, Kramers PGN, editors. *Volksgezondheid Toekomst Verkenning 1997.III Gezondheid en levensverwachting gewogen.* Maarssen: Elsevier/ De Tijdstroom; 1997.
19. *Victorian Burden of Disease Study: Morbidity.* Melbourne, Victoria: Victorian Government Department of Human Services; 1999.
20. Hölzel D, Klamers A, Schmidt M. *Krebs: Häufigkeiten, Befunde und Behandlungsergebnisse; Persepektiven für die Krebsregisterdiskussion und eine quantitative klinisch-epidemiologische Onkologie aus dem Tumorregister München.* München: W. Zuckschwerdt Verlag; 1996.
21. Reidpath DD, Allotey P. Infant mortality rate as an indicator of population health. *J Epidemiol Community Health* 2003;57(5):344-6.

Summary

Summary measures of population health combine information on mortality with information on non-fatal health outcomes. Over the past decades, the interest in and use of these measures has increased, and different types have been constructed. One distinction that can be made is between summary measures that use disease-specific epidemiological input and those that are based on generic morbidity and mortality data. As health policy decisions are mostly taken at the level of a disease or its risk factors, disease-specific summary measures of population health are in theory more useful than generic ones. However, in practice their use may be limited by their extensive data requirements. This thesis focuses on two practical data problems that may restrict the construction of disease-specific summary measures.

Firstly, disease-specific epidemiological frequency data are needed, which are not always readily available and can be of dubious quality. To remedy some of these data problems, disease models such as incidence-prevalence-mortality (IPM) models have been constructed. However, their validity and usefulness has not been studied yet. Secondly, disease-specific health status valuations (or disability weights, DWs) are required that match the severity level of the epidemiological data. Dividing diseases into stages facilitates adapting the DWs to the epidemiology, but introduces the problem of combining the stage-specific DWs into a disease-specific one. Two specific research questions were investigated:

1) To what extent are the causal disease models valid and useful to check for the consistency of epidemiological frequency data and to supplement them?
2) How can disease-stage specific health status valuations be tailored to epidemiological frequency data?

These questions were explored on the basis of empirical data for breast cancer and Major Depression (MD) in the Netherlands, both important health problems that allow us to study the research questions from different perspectives.

Part A of this thesis addresses the first research question. In chapter two we studied the validity and usefulness of IPM models, using data sets for different types of cancer. The validity of the model was supported by a check against complete and consistent artificial data. However, high-quality empirical data could not be replicated well. There is evidence that this is not only caused by data inaccuracies, but also by past trends in incidence or mortality. Although trends can be accounted for in a dynamic model, this requires additional input

data on the nature and size of trends, which are often not available. Thus, we often cannot distinguish between the effects of data problems and trends, complicating the use of IPM models in improving disease estimates.

In chapter three we compared the extent to which trends and data problems affected the model outcome for breast cancer. Two factors had a major effect: an underestimation of the 1993 prevalence data from the Regional Cancer Centre South (IKZ), and a trend of increasing incidence. Other known data problems had relatively smaller effects. Furthermore, part of the difference between model and data remained unexplained. We concluded that IPM models can be useful for checking data inconsistencies and supplementing incomplete data, but in both cases there remains the need for careful interpretation of the results: unknown data problems and trends may affect the outcome in unknown directions. In the all too common situation where, unlike for breast cancer, no data on the size and nature of trends are available, expert opinion is indispensable to assess whether models improve data quality or, inappropriately, remove the effect of trends.

In chapter four we estimated lifetime prevalence of major depression (MD) for the Netherlands and Australia from current prevalence data using a microsimulation model, and quantified the underestimation in the empirical estimate for the Netherlands. For both countries, we found that around 30% of men and 40% of women suffer from one or more episodes during their life. Lifetime rates reported from cross-sectional surveys are much lower for two reasons: 1) they ignore cases that become incident after the survey, and 2) self-reported symptoms and their timing may be biased by recall problems. Recall bias was shown to underestimate the Dutch empirical estimate by at least 35%. Modelling allowed an indirect estimation of the lifetime prevalence of depression, which is useful in the absence of direct measurement, but also indicates that the direct estimate is severely underestimated.

Part B of this thesis concerns the tailoring of the DWs. In chapter five we examined the associations of depression "severity" and type (a single or recurrent episode) with disability in the Dutch general population. Higher "severity" classes were associated with more disability, but "moderate" and "severe" MD were not significantly different. The difference between non-depressed and "mild" MD had a larger effect size than between successive "severity" classes. Recurrent MD was not associated with more disability than

single episode MD. We concluded that three clusters of MD can be used to describe the distribution of disability in the depressed population: "mild", "moderate to severe" and "severe with psychotic features". A simplification of the DSM "severity" classes could be considered, while the marked difference between "mild" and no MD suggests that "mild" cases should be considered relevant.

We studied whether disability of MD was overestimated in previous burden of disease calculations in chapter six. Previously used DWs seem to reflect a clinical situation rather than the population one from which the epidemiological data are derived. We derived DWs for the three disability clusters of MD ("mild", "moderate to severe" and "severe with psychotic features") and calculated its burden. An overall DW was based on the proportional prevalence of the clusters in the Dutch population. Our new DWs were mostly similar to previous studies, providing no indication that DWs and burden were overestimated in the past. One exception was the lower overall DW applied in the 1994 Dutch Burden of Disease Study, which was based on a proportional distribution based on the opinion of experts, who estimated a much higher proportion of cases to be "mild" than were detected in a survey of the Dutch population. This research shows the importance of the epidemiological frequency data and supports previous high estimates of the burden of MD. Finally, its results suggest that health status descriptions have only little effect on the valuation, and the disease-stage label is much more important.

In chapter seven the burden of breast cancer was compared for six European countries and its sensitivity to different sources of variation examined. The major part of the burden (more than 70%) was caused by premature mortality and the largest burdens were found in Denmark and the Netherlands, countries with the highest mortality rates. Cross-national variation in disease epidemiology was the largest source of variation, while variation in the disability weights, in demography and uncertainty in epidemiological data had smaller effects. To compare the burden of breast cancer in different countries a simple inspection of the mortality rates would have sufficed. We believe this conclusion can be extrapolated to many other diseases, as many have a large mortality component. Using country-specific DWs will add little information to the comparison of the burden for these diseases.

Chapter eight, the general discussion, integrates and discusses the results from the six studies. The research showed that disease models can be used to detect consistency problems in the epidemiological data and to supplement them, although trends complicate their use. Furthermore, different ways were found to tailor stage-specific health status valuations to the epidemiology. On these grounds we conclude that useful disease-specific summary measures can be constructed in practice. Nevertheless, disease-specific summary measures should not be interpreted as being very precise and their practical application is restricted by the abundance of diseases.

The results of the studies stressed the importance of obtaining sound epidemiological data for the calculation of disease-specific summary measures. Research should therefore focus on obtaining consistent epidemiological data, supplementing empirical data collection by modelling exercises. The DWs on the other hand, were shown to be often of less influence on the final summary measure. Efforts to improve them should focus mainly on combining the stage-specific weights into an overall disease-specific one. Finally, the abundance of diseases will restrict the application of summary measures. Therefore the intended use of a measure should guide the choice between a generic or disease-specific measure, and the design of the measure in general.

Summary

in Dutch

Samengestelde volksgezondheidsmaten vatten gegevens over sterfte en ziekte samen in één getal. In de afgelopen decennia is de belangstelling voor en het gebruik van deze maten sterk toegenomen en zijn er verschillende varianten ontworpen. We onderscheiden onder andere samengestelde volksgezondheidsmaten die worden berekend op basis van ziektespecifieke epidemiologische gegevens en maten die gebruik maken van generieke informatie over sterfte en ziekte. Aangezien beleidsbeslissingen in de gezondheidszorg meestal gericht zijn op specifieke ziekten of risicofactoren, zijn ziektespecifieke samengestelde volksgezondheidsmaten beter bruikbaar dan generieke, tenminste in theorie. In de praktijk wordt het gebruik van deze maten beperkt door de vele gegevens die nodig zijn om ze te berekenen. Dit proefschrift richt zich op twee praktische data problemen die de constructie en het gebruik van samengestelde volksgezondheidsmaten kunnen belemmeren.

Ten eerste zijn ziektespecifieke epidemiologische frequentie gegevens nodig. Deze zijn niet altijd voorhanden, en zijn bovendien vaak van twijfelachtige kwaliteit. Om dit soort data problemen te verhelpen zijn ziekte modellen zoals incidentie-prevalentie-mortaliteit (IPM) modellen ontworpen. De validiteit en bruikbaarheid van deze modellen zijn nog niet onderzocht. Ten tweede zijn ziektespecifieke waarderingen voor gezondheidstoestanden (ofwel: wegingsfactoren) nodig, die precies aansluiten op de epidemiologische data. Het verdelen van ziekten in ziektestadia maakt het makkelijker om wegingsfactoren op de epidemiologische gegevens te laten aansluiten, maar veroorzaakt een nieuw probleem: hoe deze stadiumspecifieke wegingsfactoren te combineren tot één ziektespecifiek gewicht? In dit proefschrift worden twee onderzoeksvragen bestudeerd:

1) In hoeverre zijn ziektemodellen valide en bruikbaar om de consistentie van epidemiologische gegevens te testen en om deze gegevens aan te vullen?
2) Hoe kunnen waarderingen voor de gezondheidstoestand van patiënten in specifieke ziektestadia worden afgestemd op epidemiologische frequentie gegevens?

Deze vragen werden bestudeerd aan de hand van empirische data voor borstkanker en depressie in Nederland. Beide zijn belangrijke gezondheidsproblemen die door hun verschillen het mogelijk maken de onderzoeksvragen vanuit verschillende perspectieven te bestuderen.

Deel A van dit proefschrift richt zich op de eerste onderzoeksvraag. In hoofdstuk twee bestudeerden we de validiteit en bruikbaarheid van IPM modellen, door gebruik te maken van gegevens voor verschillende soorten kanker. De validiteit van het model werd bevestigd door het te testen met artificiële gegevens die per definitie volledig en consistent zijn. Daarentegen kon het model kwalitatief goede empirische gegevens niet reproduceren. Dit wordt niet alleen wordt veroorzaakt door inconsistenties in de gegevens, maar ook door trends over tijd in incidentie of sterfte. Alhoewel in een dynamisch model rekening gehouden kan worden met tijdstrends, zijn daarvoor extra gegevens nodig over de aard en grootte van de trend. Deze zijn vaak niet beschikbaar. Voor veel aandoeningen kan er dus in het model geen onderscheid gemaakt worden tussen effecten van data problemen en van trends. Dit compliceert het gebruik van IPM modellen om gegevens te verbeteren.

In hoofdstuk drie vergeleken we in hoeverre tijdstrends en data problemen de modelberekeningen voor borstkanker beïnvloeden. Twee factoren hadden een groot effect: een onderschatting van de borstkanker prevalentie en een stijging van de incidentie in de onderzochte periode. Andere data problemen die bekend waren hadden een relatief kleinere invloed. Daarnaast kon een deel van het verschil tussen modelberekeningen en gegevens niet verklaard worden. We concludeerden dat IPM modellen nuttig kunnen zijn om inconsistenties in gegevens op te sporen en incomplete gegevens aan te vullen, maar dat de modelberekeningen voorzichtig moeten worden geïnterpreteerd: onbekende data problemen en trends over de tijd kunnen de uitkomsten in onbekende richting beïnvloeden. In de meeste gevallen zijn geen gegevens beschikbaar over de aard en grootte van trends. In die gevallen is het oordeel van experts onmisbaar om te beoordelen of het model, zoals gewenst, de kwaliteit van de data verbetert of, ten onrechte, het effect van trends corrigeert.

In hoofdstuk vier schatten we met behulp van een microsimulatie model de *lifetime* prevalentie (ooit depressie gehad) van depressie in Nederland en Australië. Voor beide landen vonden we dat ongeveer 30% van de mannen en 40% van de vrouwen ooit een episode van depressie hebben gehad. De percentages die in cross-sectionele studies worden gerapporteerd zijn veel lager om twee redenen: 1) mensen die een depressie krijgen na het onderzoek worden niet meegeteld, 2) zelfrapportage van symptomen en het moment waarop die aanwezig waren kan door herinneringsproblemen aan bias onderhevig zijn. De berekeningen toonden aan dat de empirische schatting van lifetime prevalentie

in Nederland met ten minste 35% onderschat wordt ten gevolge van deze "herinnerings-bias". Dit zet vraagtekens bij de empirische meting van lifetime prevalentie. Het modelleren maakte het mogelijk om een indirecte schatting van de lifetime prevalentie van depressie te verkrijgen. Dit is waardevol wanneer empirische gegevens uit langdurige follow-up studies ontbreken.

Deel B van dit proefschrift heeft betrekking op het afstemmen van de wegingsfactoren op de epidemiologie. In hoofdstuk vijf onderzochten we de associaties van de ernst van depressie en het type depressie (eenmalige en recidiverende depressie) met functionele beperkingen in de Nederlandse algemene bevolking. Hogere ernstklassen waren geassocieerd met meer beperkingen, zij het dat "matige" en "ernstige" depressie niet significant van elkaar verschilden. Het verschil tussen niet-depressief en "licht" depressief was groter dan verschillen tussen de opvolgende ernstklassen. Recidiverende depressie was niet geassocieerd met meer functionele beperkingen dan eenmalige depressie. We concludeerden dat drie clusters van depressie gebruikt kunnen worden om de verdeling van beperkingen in de depressieve populatie te beschrijven: "lichte", "matige tot ernstige" en "ernstige depressie met psychotische kenmerken". Een vereenvoudiging van de DSM ernst categorieën zou kunnen worden overwogen. Het aanzienlijke verschil in functionele beperkingen tussen "licht" depressief en niet depressief suggereert dat "lichte" depressie als relevant dient te worden beschouwd.

Of de ziektelast van depressie werd overschat in eerdere ziektelast berekeningen bestudeerden we in hoofdstuk zes. Eerder gebruikte wegingsfactoren voor depressie lijken meer de klinische situatie te weerspiegelen dan de situatie in de populatie, waaruit de epidemiologische gegevens zijn afgeleid. We leidden wegingsfactoren af voor de drie clusters van beperkingen voor depressie ("licht", "matig tot ernstig" en "ernstig met psychotische kenmerken") en berekenden de ziektelast van depressie. Een gemiddelde wegingsfactor voor depressie werd berekend op basis van de proportionele prevalenties van de drie clusters in de Nederlandse bevolking. De nieuwe wegingsfactoren waren over het algemeen gelijk aan eerdere studies, zodat er geen aanleiding was om te veronderstellen dat de ziektelast in eerdere studies werd overschat. Eén uitzondering hierop was de gemiddelde wegingsfactor uit de Nederlandse Ziektelast Studie 1994. Deze was gebaseerd op expert schattingen van de proportionele prevalenties van de clusters. Het

aandeel van "lichte" depressie was in deze schattingen veel hoger dan in een steekproef van de Nederlandse bevolking. Dit onderzoek laat zien hoe belangrijk de epidemiologische frequentie gegevens zijn in ziektelast berekeningen en ondersteunt eerdere hoge schattingen van de ziektelast van depressie.

In hoofdstuk zeven vergeleken we de ziektelast van borstkanker in zes Europese landen en onderzochten we de gevoeligheid van de schatting van de ziektelast voor verschillende bronnen van variatie. Het grootste deel van de ziektelast (meer dan 70%) werd veroorzaakt door vroegtijdige sterfte. De grootste ziektelast werd dan ook gevonden in Denemarken en Nederland: de landen met de grootste sterfte aan borstkanker. Variatie in de epidemiologische frequentie gegevens tussen landen was de grootste bron van variatie, terwijl variatie in wegingsfactoren, in demografie, en de onzekerheid in de epidemiologische gegevens kleinere effecten hadden. Hieruit concludeerden we dat om de ziektelast van borstkanker in de verschillende landen te vergelijken een simpele inspectie van de sterfte gegevens voldoende was. We geloven dat dit ook voor veel andere ziekten zal gelden, aangezien veel ziekten een hoge sterfte component hebben. Het gebruik van verschillende wegingsfactoren per land zal voor die ziekten weinig informatie toevoegen aan de vergelijking van de ziektelast.

Hoofdstuk acht integreert en bespreekt de resultaten van deze zes studies. Het onderzoek laat zien dat ziektemodellen gebruikt kunnen worden om problemen in epidemiologische gegevens op te sporen en om deze aan te vullen, alhoewel tijdstrends hun gebruik bemoeilijken. Daarnaast werden verschillende manieren gevonden om stadiumspecifieke wegingsfactoren aan te laten sluiten op de epidemiologie. Op deze gronden concluderen we dat het in de praktijk vaak mogelijk is om bruikbare ziektespecifieke samengestelde volksgezondheidsmaten te berekenen. Ziektespecifieke maten zijn echter niet erg nauwkeurig en hun praktische toepassing voor de beschrijving en vergelijking van de totale volksgezondheid in populaties wordt beperkt door de grote hoeveelheid ziekten.

De resultaten van de studies benadrukken het belang van het verzamelen van betrouwbare epidemiologische gegevens voor het berekenen van ziektespecifieke samengestelde volksgezondheidsmaten. Onderzoek moet zich richten op het verkrijgen van consistente epidemiologische data, door empirische

data verzameling, en aanvullend daarop het gebruik van modellen. Ook werd aangetoond dat wegingsfactoren minder invloed op de uiteindelijke samengestelde maat hebben. Inspanningen om wegingsfactoren te verbeteren zouden zich met name moeten richten op het combineren van stadiumspecifieke wegingsfactoren.

Tot slot, samengestelde volksgezondheidsmaten die zijn opgebouwd uit alle in een populatie voorkomende relevante ziekten, stellen extreem hoge eisen aan de beschikbaarheid en betrouwbaarheid van de epidemiologische gegevens. Het beoogde gebruik van een maat moet bepalend zijn voor het aantal op te nemen ziekten en de mate van detaillering, evenals voor het verdere ontwerp ervan. Het ontwikkelen van samengestelde volksgezondheidsmaten blijft "maatwerk", waarbij het uiteindelijke doel en de beschikbaarheid van gegevens maatgevend zijn.

Word of thanks

I would like to thank all those who have contributed to this thesis. In the first place my many supervisors, who have helped to bring this project to fruition: Professor Paul van der Maas, Professor Daan Kromhout, Dr. Jan J. Barendregt, Dr. Nancy Hoeymans, Dr. Marie-Louise Essink-Bot en dr. Pieter G.N. Kramers. Working at two institutes and with as many as six different supervisors was a privilege. I have learned a lot from you and I am very grateful of having all of you as supervisors.

Dear Paul, many thanks for your supervision, constructive comments and advises both at meetings and in writing, and for your trust. Dear Daan, I would like to thank you for your critical look and valuable comments to this thesis. Dear Jan, your door was always open. Thank you for introducing me into the world of IPM models, for your supervision and guidance, and for getting me back on track whenever I was stuck. Dear Nancy, you were always there when I had questions, needed help or advice, and last but not least encouragement. You helped me through my first real statistical analysis, and made sure I kept having faith in this project. Dear Marie-Louise, you are my virtual third co-promotor, as you have contributed not only to at least half of the studies reported in this thesis, but also to the direction the project as a whole was taking. Thank you for your thorough look both at the design of studies as well as the manuscripts (your "stofkam"), helping me to organise them, and to get through the BIOMED project. Pieter, I greatly appreciate your advice and comments to several manuscripts and to the start-up of the project as a whole.

I am grateful to have had so many great colleagues both at the Department of Public Health (MGZ) as at the Department of Public Health Forecasting (VTV). I have always felt very much at home at both departments. I would like thank the VTV team for the great ambiance they make, and their readiness to help: you are a great team! In particular I would like to thank Andrea, whom I have shared a room with for more than two years, for our valuable conversations and good times! At MGZ I am obliged to many colleagues without whom the department wouldn't be the pleasant and inspiring workplace that it is. Especially Anna, Jetty and Oscar who have helped me on many occasions: thanks!

Also, I would like to thank the team at the department of human services in Melbourne for my great time there: Theo, Anne, Michelle, Sunil, Barbara, Michael, and the rest, thank you ever so much! Theo, thank you for supervising me in the depression modelling work, and in boogy-boarding! Anne, thank you for your friendship, making me feel at home, giving me a home, and helping me with just about everything!

Furthermore, many thanks to Oscar and Daan Vernooy, for checking this manuscript. Finally, a word of thanks to Oz, my parents, and all my friends and family, for making life a great experience!

Curriculum Vitae

Michelle Kruijshaar was born on November 3, 1972 in Amsterdam, the Netherlands. She graduated from secondary school in 1991 at the "Stedelijk Gymnasium" in Leiden. After working as a care-assistant in Newport, South-Wales (U.K.), and studying one year at the "Vrije Hogeschool" in Driebergen, she studied medical biology at the "Vrije Universiteit" in Amsterdam from 1993 until 1998.

In February 1999 she started as a PhD student at the Department of Public Health of the Erasmus MC in Rotterdam and the Department for Public Health Forecasting of the National Institute for Public Health and the Environment in Bilthoven. In 2001 she obtained a master's degree in epidemiology at the Netherlands Institute for Health Sciences (nihes) (2001). She made a working visit to Australia, where she worked with dr. T. Vos at the Department of Human Services, Melbourne, and visited professor G. Andrews at the Clinical Research Unit on Anxiety and Depression of the University of New South Wales, Sydney (November 2002 to January 2003).

Since August 2003 she has been working as a postdoc at the department of Public Health, on the description an evaluation of the burden of diagnostic interventions.

Publications

Michelle E. Kruijshaar, Jan J. Barendregt, Nancy Hoeymans. The use of models in the estimation of disease epidemiology. *Bulletin of the World Health Organization* 2002; 80 (8): 622-628.

Michelle E. Kruijshaar, Jan J. Barendregt, Lonneke V. van de Poll-Franse. Estimating the prevalence of breast cancer using a disease model: data problems and trends. *Population health metrics* 2003; 1 (1): 5.

Michelle E. Kruijshaar, N. Hoeymans, R.V. Bijl, J. Spijker, M.L. Essink-Bot. Levels of disability in Major Depression. Findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Journal of Affective Disorders* 2003; 77 (1): 53-64.

Michelle E. Kruijshaar, Jan J. Barendregt, and the European Disability Weights Group. The breast cancer related burden of morbidity and mortality in six European countries: The European Disability Weights projects. *The European Journal of Public Health*, in press.

Michelle E. Kruijshaar, Jan J. Barendregt, Theo Vos, Ron de Graaf , Jan Spijker, Gavin Andrews. Lifetime prevalence estimates of major depression: an indirect estimation method and a quantification of recall bias. *Submitted.*

Michelle E. Kruijshaar, Nancy Hoeymans, Jan Spijker, Marlies E.A. Stouthard, Marie-Louise Essink-Bot. The burden of depression in the Netherlands: Is disability overestimated? *Submitted.*

Jan J. Barendregt & Michelle E. Kruijshaar. An indirect estimate of episode duration in major depression. *Submitted.*

Theo Vos, Michelle M Haby, Jan J Barendregt, Michelle E. Kruyshaar, Justine Corry, Gavin Andrews. The burden of major depression avoidable by alternative treatment strategies. *Submitted.*

Publication of this thesis was financially supported by the Department of Public Health, University Medical Center Rotterdam and the National Institute of Public Health and the Environment, Bilthoven.

The research described in this thesis was partly prepared at the Department of Public Health, University Medical Center Rotterdam and the National Institute of Public Health and the Environment, Bilthoven.